# A Novel Method for Assessing Task Complexity in Outpatient Clinical-Performance Measures

Sylvia J. Hysong, Ph.D.[1,2], Amber B. Amspoker, Ph.D[1,2], and Laura A. Petersen, M.D., M.P.H[1,2]

[1]Center for Innovations in Quality, Safety, and Effectiveness, Michael E. DeBakey VA Medical Center, Houston, TX, USA; [2]Department of Medicine – Health Services Research Section, Baylor College of Medicine, Houston, TX, USA.

***BACKGROUND:*** Clinical-performance measurement has helped improve the quality of health-care; yet success in attaining high levels of quality across multiple domains simultaneously still varies considerably. Although many sources of variability in care quality have been studied, the difficulty required to complete the clinical work itself has received little attention.

***OBJECTIVE:*** We present a task-based methodology for evaluating the difficulty of clinical-performance measures (CPMs) by assessing the complexity of their component requisite tasks.

***DESIGN:*** Using Functional Job Analysis (FJA), subject-matter experts (SMEs) generated task lists for 17 CPMs; task lists were rated on ten dimensions of complexity, and then aggregated into difficulty composites.

***PARTICIPANTS:*** Eleven outpatient work SMEs; 133 VA Medical Centers nationwide.

***MAIN MEASURES:*** *Clinical Performance*: 17 outpatient CPMs (2000–2008) at 133 VA Medical Centers nationwide. *Measure Difficulty*: for each CPM, the number of component requisite tasks and the average rating across ten FJA complexity scales for the set of tasks comprising the measure.

***KEY RESULTS:*** Measures varied considerably in the number of component tasks ($M = 10.56$, $SD = 6.25$, min = 5, max = 25). Measures of chronic care following acute myocardial infarction exhibited significantly higher measure difficulty ratings compared to diabetes or screening measures, but not to immunization measures ($\bar{z} = 0.45$, −0.04, −0.05, and −0.06 respectively; $F_{(3,\ 186)} = 3.57$, $p = 0.015$). Measure difficulty ratings were not significantly correlated with the number of component tasks ($r = -0.30$, $p = 0.23$).

***CONCLUSIONS:*** Evaluating the difficulty of achieving recommended CPM performance levels requires more than simply counting the tasks involved; using FJA to assess the complexity of CPMs' component tasks presents an alternate means of assessing the difficulty of primary-care CPMs and accounting for performance variation among measures and performers. This in turn could be used in designing performance reward programs, or to match workflow to clinician time and effort.

*KEY WORDS:* clinical performance measurement; outpatient; task complexity; functional job analysis.

***Electronic supplementary material*** *The online version of this article (doi:10.1007/s11606-015-3568-z) contains supplementary material, which is available to authorized users.*

*Published online March 7, 2016*

J Gen Intern Med 31(Suppl 1):S28–35
DOI: 10.1007/s11606-015-3568-z


## INTRODUCTION

According to the Institute of Medicine and others, the quality of health-care delivery in the United States has improved markedly in the last decade, partly due to the proliferation of clinical-performance measures used by health-care organizations to assess and monitor their quality and make improvements accordingly, though much variability still exists.[1,2]

In health care, performance measurement systems are commonly criticized for not adjusting for patient complexity, purportedly important because sicker, more complex patients take more effort to treat. Some efforts to capture this construct have been noted, such as using Diagnostic Related Group (DRG) codes to measure the degree of hospital resources needed to treat a given diagnosis, or risk adjusting clinical performance by the number of comorbidities associated with a patient. However, clinical performance measures assess specific processes that go into the care of a clinical condition. Evidence that performance measures tend to cluster by process more strongly than by disease[3] would suggest that disease severity is inadequate to reliably and validly measure the level of effort needed to successfully satisfy the criteria outlined in each performance measure. For example, the same screening tool is used to determine whether a patient has mild or severe depression; thus, the same level of effort is exerted to successfully execute the tasks that satisfy the performance measure (screen the patient for depression), regardless of the patient's disease severity.

Outside healthcare, a key method of characterizing work is by measuring the complexity and difficulty of the tasks involved. Task complexity and difficulty are central to establishing performance measurement criteria,[4] and have long been known to relate to task performance[5,6]: the more components, inputs, and relationships a task possesses (task complexity), the more effort is needed (task difficulty) for a person with a given level of ability to successfully perform the task.[7] Within

health care, previous studies have shown variability in the complexity and difficulty of tasks in primary care settings.[8,9] This suggests that, depending on which tasks are performed to meet the standards required by a clinical-performance measure, different performance measures are likely to exhibit different degrees of difficulty. Performance-measure difficulty is defined as the effort required to perform the set of tasks comprising a clinical performance measure, such that the standard required by the measure is successfully achieved. In other words, we posit that clinical-performance measure standards are more or less difficult to reach due to the difficulty of the measures' component tasks. As a result, we can evaluate a clinical performance measure's *difficulty* as a potential variability source in measured quality of care. The relationships among task complexity, task difficulty, and measure difficulty are summarized in Box 1.

Box 1. Understanding the relationships between complexity and difficulty of tasks and measures

- ***Task complexity***: An invariant property of a task characterized by the number of components, inputs, products, and the relationships among all three involved in performing a task. In this paper, task complexity is operationalized by individual task ratings of 10 dimensions recommended by the Functional Job Analysis (FJA) methodology.
- ***Task difficulty***: the level or degree of effort needed by an individual to successfully perform the task – operationalized here as the average of all the complexity ratings of a single task (please note: tasks are by definition performed by a single individual, whose ability may facilitate or hinder their success at exerting the effort needed to successfully perform the task)
- ***Measure difficulty***: degree of effort required to perform the set of tasks comprising a clinical performance measure, such that the standard required by the measure is successfully achieved—operationalized here as the average of all task difficulty ratings for the set of tasks comprising a given measure.
- ***Measure complexity***: for the set of tasks comprising a given measure, measure complexity is the average of all task complexity ratings on a given FJA scale. Each clinical-performance measure has one measure-complexity score for each of the ten FJA scales.

The table below visually depicts the relationships among the concepts described in Box 1.

| | Scale 1 (e.g., Things) | Scale 2 (e.g., Data) | Scale J | Average (Scale 1..J) |
|---|---|---|---|---|
| Measure K | | | | |
| Task 1 | Task 1 complexity (with respect to Scale 1) | Task 1 complexity (with respect to Scale 2) | Task 1 complexity (with respect to Scale J) | Task 1 Difficulty |
| Task 2 | Task 2 complexity (with respect to Scale 1) | Task 2 complexity (with respect to Scale 2) | Task 2 complexity (with respect to Scale J) | Task 2 Difficulty |
| Task n | Task n complexity (with respect to Scale 1) | Task n complexity (with respect to Scale 2) | Task n complexity (with respect to Scale J) | Task n Difficulty |
| Average (task 1..n) | Measure K complexity (with respect to Scale 1) | Measure K complexity (with respect to Scale 2) | Measure K complexity (with respect to Scale J) | Measure K Difficulty |

Many sources of variability in care quality have been studied and documented, including patient factors such as disease burden and comorbidities[10,11]; organizational factors such as availability of resources, size, academic mission, and financial incentives[12]; and provider factors such as knowledge/training and abilities. All these have been shown to impact quality of health care to various degrees. These effects are consistent with Work-Doing Systems theory,[13] which proposes a dynamic interaction of three basic forces influencing performance: the work organization (its purpose, goals, and resources); the worker (individual skills and abilities, experience, education and training); and the work itself (tasks and associated performance standards). Though much research exists linking the work organization and the worker to clinical performance,[12,14–16] the work itself has received the least attention in health-care quality research.

For example, at the Department of Veterans Affairs, meeting the tobacco-cessation performance measure requires little more than providing and documenting brief counseling of the patient via a checkbox in the electronic health record. In contrast, a blood-pressure-control measure may require numerous multicomponent tasks, including monitoring patient blood pressure, adjusting medications as needed, advising the patient regarding lifestyle modifications, and providing patient education to ensure adherence with the treatment plan. The differences in difficulty of the measure's component tasks suggest that, all other things being equal, it should be easier to attain higher quality levels for the tobacco-cessation measure (as currently defined) than for the blood-pressure-control measure.

The difficulty of a performance measure is important to measure accurately, as it can be used to make numerous important administrative decisions. For example, difficulty can be used as a risk-adjustment variable when creating performance composites for providers and facilities, or to help select measures worth monitoring (measures that are too easy or too difficult to achieve may be not worth tracking because they provide no variance). It could also be built into reward systems (measures that are more difficult could be rewarded more than easier ones).

The concept of difficulty or effort is addressed to some extent in the payment systems literature with the advent of Resource-Based Relative Value Units,[17] which considers "total work input by the physician" as one of three factors describing the resource-based relative value of a given medical service. RRVUs are operationalized as a combination of four dimensions : (1) time; (2) mental effort and judgment; (3) technical skill and physical effort; and (4) psychological stress. Such systems conflate characteristics of the work (e.g., mental effort, physical effort) with characteristics of the worker (e.g., technical skill, judgment). This is likely of little consequence when the purpose of the system is physician billing. If, the goal is to understand what improves quality, regardless of who does the work, a more nuanced approach is required.

We offer a methodology for characterizing the difficulty of accomplishing clinical performance measures that begins with rating each task required to complete the measure across ten domains of complexity, and then aggregates scores for each measure. The result is defined as measure difficulty. We hypothesized that this method would result in significant variation in difficulty among clinical-performance measures.

## METHODS

### Design

Local Institutional Review Board approval was obtained for the study. We employed Functional Job Analysis (FJA), a methodology from industrial/organizational psychology based on Work-Doing Systems Theory [13,18,19] that is used for describing and assessing work complexity. FJA was originally used by the U.S. Department of Labor in developing its *Dictionary of Occupational Titles* (the standard for developing most job descriptions) and its current electronic counterpart, O*NET.[20,21] In healthcare, it has been used to describe the complexity of primary care and in reallocating work among clinicians and systems redesign.[8,9,22] It is particularly suited for this study, because its basic unit of analysis is the task, the smallest, naturally self-contained unit of work performed by an individual to accomplish a specific result. The task level permits nuanced understanding not achievable at the procedure/service level (as with RBRVUs), without micro-level, time-and-motion style analyses. Additionally, tasks are rated behaviorally (and thus concretely) using Guttman scales that allow specific behavioral meanings to be ascribed at any given level on the scale.

We sought to quantify the difficulty of all tasks required to accomplish 17 outpatient clinical-performance measures used in the Veterans Health Administration (VHA), described below. Appendix A (available online) presents a primer on the FJA methodology; highlights as they relate to our study are summarized below.

### Clinical Performance Measure Selection

We selected 17 outpatient clinical measures from the VHA's External Peer Review Program (EPRP) to assess clinical performance. EPRP is one of the official data sources for VHA's clinical-performance-management system and is used by leadership to make administrative decisions about facilities. Beginning in fiscal year (FY) 2000 and updated quarterly, EPRP abstracts clinical data directly from electronic medical records to calculate inpatient and outpatient performance measures for all VA Medical Centers nationally. Measures were selected using the following criteria: (a) outpatient care only; (b) method of calculation has remained unchanged for at least 4 years; and (c) focus on chronic/preventive processes or intermediate outcomes only. Appendix B (available online) lists the measures selected, their technical definitions, and the set of component tasks required to accomplish each.

### Participants

Eight primary care physicians from two geographically dispersed VA Medical Centers served as subject-matter experts to help develop the list of tasks corresponding to each clinical-performance measure.[23,24] This sample size is consistent with standard recommendations for FJA and focus group methologies.[13,25] Experts were selected for their experience with clinical-performance data and intimate knowledge of the clinical processes involved.

### Procedure

***Identifying the Tasks Required for Each Clinical-Performance Measure.*** We used FJA to create standardized lists of the specific tasks required to successfully perform each clinical quality measure to its required standard.[13,26] Appendix A (available online) presents an FJA primer, a sample clinical-performance measure, and its accompanying task set.

The experts reviewed the definition of each clinical-performance measure[27,28] and independently listed all tasks required to satisfy its performance standard. The research team compiled experts' responses into a single task list for each measure. Tasks were cross-checked against a validated FJA-compliant primary care task database.[29] Valid and reliable task complexity ratings (described below) require a consistent content and linguistic structure from task to task.[26] For tasks with matching database tasks, we used the database task. A one-half–day focus group with the experts was employed for tasks without a database match to ensure a consistent, FJA-compliant structure. For each unmatched task, experts described the specific actions, knowledge, skills, abilities, and tools required to achieve the result listed in the task. Our primary care subject-matter experts (SMEs) were unable to describe some tasks (e.g., conducting a colonoscopy) to FJA specifications; in such cases, the SMEs referred us to the appropriate experts (a gastroenterologist, gynecologist, and lab manager) with whom we repeated the procedure. The research team edited the lists to ensure adherence to FJA structure. The SMEs reviewed the final task sets to ensure that each comprised at least 85 % of the work required to meet the relevant measure.[13]

***Rating Task Complexity.*** Two trained research teammembers, one of whom also served as the focus group facilitator, independently rated each newly generated task statement on ten complexity dimensions using the scales prescribed by FJA (see Table 1 for brief definitions and scale ranges; see Appendix A (available online) for more detail). Using trained raters instead of subject-matter experts ensured the ratings were based exclusively on the language in the written task statement and scale definitions, rather than an expert's general knowledge or schema, which could contaminate the validity of the complexity ratings.[26] Rating discrepancies were resolved by consensus.

**Table 1 Functional Job Analysis Rating Scales and Brief Definitions**

| Scale | Range of scores* | | | Definition |
|---|---|---|---|---|
| | Low | Med | High | |
| Things | 1–2 | 3 | 4 | Physical interaction with and response to tangibles—touched, felt, observed, and related to in space; images visualized spatially |
| Data | 1–2 | 3–4 | 5–6 | Information, ideas, facts, statistics, specification of output, knowledge of conditions, techniques; mental operations |
| People | 1–2 | 3–4 | 5–8 | Live interaction among people, and between people and animals |
| Worker Instructions | 1–2 | 3–4 | 5–8 | Amount of autonomy afforded worker, based on the degree to which inputs, outputs, tools, and procedures required to accomplish task are specified |
| Reasoning | 1–2 | 3–4 | 5–6 | Knowledge, ability to deal with theory versus practice, abstract versus concrete and many versus few variables |
| Mathematics | 1–2 | 3 | 4–5 | Knowledge and ability to deal with mathematical problems and operations from counting and simple addition to higher mathematics |
| Language | 1–2 | 3–4 | 5–6 | Knowledge and ability to speak, read, or write language materials from simple verbal instructions to complex sources or written information and ideas |
| Worker Technology | 1–2 | 3–4 | 5–8 | Means and methods employed in completing a task or work assignment (tools, machines, equipment or work procedures, processes or any other aids to assist in the handling, processing or evaluation of things or data |
| Worker Interaction | 1–2 | 3–4 | 5–8 | Degree to which, when working with others (through direct or indirect contact), workers assist each other, coordinate their efforts and adapt their style and behavior to accommodate atypical or unusual circumstances and conditions; this effort leads to achieving employer goals to given standards |
| Human-Error Consequence | 1–2 | 3–4 | 5–8 | Degree of responsibility imposed upon the performer with respect to possible mental or physical harm to persons (including performer, recipients, respondents, co-workers, or the public), resulting from errors in performance of the task being scaled |

*1 is the lowest level of complexity associated with each scale, representing the lowest behavioral benchmark for the scale in question. Higher numbers mean higher degrees of complexity on each given scale. Each scale has a different maximum, because the scales are benchmarked to their natural behavioral limits. For example, complexity with respect to data was benchmarked on six naturally occurring levels: (1) comparing, (2) copying, (3) computing/compiling, (4) analyzing, (5) innovating, and (6) synthesizing. Each scale is benchmarked in a similar manner, yielding to different natural ranges. See Fine and Cronshaw[22] for benchmark levels associated with each FJA scale. Fine and Getkate[19] defined low, medium, and high ranges for the Things, Data, and People scales. Ranges on the remaining scales were grouped into low, medium, and high, accordingly*

## Data Analysis

***Calculating Performance-Measure Difficulty.*** For each clinical-performance measure, we calculated composite, measure-level complexity ratings along each FJA scale by averaging all task ratings on that particular scale. As each scale had a different possible range, we first calculated standardized (z) mean scores, then averaged these *z*-scores to arrive at a composite numerical assessment of difficulty for each measure (see Box 1). The measures selected represented four different care areas (i.e., chronic care, screening, diabetes, and immunization). We obtained a difficulty score for each care area by first standardizing scores on the ten FJA scales and then averaging the standardized scores of all clinical-performance measures belonging to that care area.

***Measure Difficulty as Number of Tasks.*** In addition to the composite calculations of measure difficulty described above, we used the number of tasks involved in each clinical-performance measure as a simple proxy for measure difficulty. Both Pearson's bivariate and Spearman's rank-order correlation coefficients were calculated to evaluate the relationship between number of component tasks in a measure and the difficulty composite, as well as the relationships between number of tasks and each of the ten scale complexity ratings.

***Differences Among Clinical-Performance Measures and Care Areas in Measure Difficulty Ratings.*** We examined differences in mean measure difficulty scores among the clinical-performance measures studied and the four specific care areas via a one-way between-groups analysis of variance (ANOVA), using Tukey's WSD to correct for multiple comparisons and the Brown-Forsyth test (with Welch's correction when needed) to check homogeneity of variance.

***Differences Among Clinical-Performance Measures in FJA Scale Complexity Ratings.*** We examined differences among the clinical measures on each of the ten FJA complexity scales using the same one-way between-groups ANOVA method described above.

# RESULTS

## Correspondence of Clinical-Performance Measures and Difficulty Scores

Our intent was to apply FJA methodology to 17 measures VHA uses to measure clinical performance in order to assess measure difficulty. However, in two instances there was not a

**Table 2 Standardized (z-score) Measure Complexity and Difficulty Ratings, and Results of Analyses of Variance across Measures/Care Areas**

| Clinical performance (EPRP) measure | *n of tasks* | Measure complexity ratings (FJA Scales) | | | | | | | | | | Measure difficulty | | Measure difficulty by care area | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Scale → | P | T | D | WI | R | M | L | WT | WINT | HEC | Mean | SD | Mean | SD |
| Care area: chronic care following AMI | | | | | | | | | | | | | | | |
| 1. Aspirin at most recent visit | 7 | 0.49 | −0.07 | 0.33 | 0.57 | 0.27 | 0.94 | 0.38 | 0.49 | 0.17 | 0.87[a] | 0.44 | 0.47 | 0.45[a] | 0.56 |
| 2. Beta blockers at most recent visit | 11 | 0.58 | 0.18 | 0.29 | 0.39 | 0.44 | 0.96[a] | 0.27 | 0.58 | 0.29 | 0.53 | 0.45[a] | 0.63 | | |
| Care area: screening | | | | | | | | | | | | | | | |
| 3. Breast cancer | 15 | −0.05 | −0.28 | −0.09 | −0.01 | −0.01 | −0.44[b] | 0.06 | 0.15 | −0.60 | −0.22 | −0.15 | 0.53 | −0.05[b] | 0.62 |
| 4. Cervical cancer | 25 | −0.49 | 0.10 | −0.34 | −0.39 | −0.50 | −0.41[b] | −0.43 | −0.42 | −0.41 | −0.62[b] | −0.39[b] | 0.59 | | |
| 5a. Colorectal cancer - colonoscopy | 24 | −0.01 | 0.24 | 0.01 | 0.01 | 0.09 | −0.05 | −0.06 | −0.15 | −0.02 | 0.44[a] | 0.05 | 0.60 | | |
| 5b. Colorectal cancer – flex. sig. | 19 | 0.00 | 0.05 | 0.09 | −0.14 | −0.01 | −0.13 | 0.07 | −0.10 | 0.05 | 0.29 | 0.02 | 0.59 | | |
| 6. Major depressive disorder | 9 | 0.42 | −0.16 | 0.48 | 0.78 | 0.59 | 0.36 | 0.51 | 0.42 | 0.59 | 0.97[a] | 0.50[a] | 0.53 | | |
| 7. Tobacco use (Past 12 Months) | 5 | 0.40 | −0.19 | 0.48 | 0.19 | 0.16 | 0.27 | 0.38 | 0.63 | 0.03 | 0.29 | 0.26 | 0.50 | | |
| Care area: diabetes mellitus / hypertension | | | | | | | | | | | | | | | |
| 8. Dx HTN and BP >160/100 or not recorded | 7 | 0.33 | 0.14 | 0.33 | 0.43 | 0.38 | 0.33 | 0.04 | 0.49 | 0.62 | −0.20 | 0.29 | 0.77 | −0.04[b] | 0.65 |
| 9. BP >160/100 or not done | 7 | 0.33 | 0.14 | 0.33 | 0.43 | 0.38 | 0.33 | 0.04 | 0.49 | 0.62 | −0.20 | 0.29 | 0.77 | | |
| 10. Foot inspection | 6 | −0.56 | 0.24 | 0.12 | −0.39 | 0.16 | −0.23 | 0.18 | 0.15 | −0.55 | −0.17 | −0.10 | 0.44 | | |
| 11. Foot-pedal pulses | 6 | −0.56 | 0.24 | 0.12 | −0.39 | 0.16 | −0.23 | 0.18 | 0.15 | −0.55 | −0.17 | −0.10 | 0.44 | | |
| 12. HbA1c >11 or none past year | 11 | 0.18 | −0.34 | −0.01 | 0.13 | 0.01 | 0.19 | −0.05 | −0.28 | 0.15 | −0.34 | −0.04 | 0.68 | | |
| 13. HbA1c <9 (Good Control) | 11 | 0.18 | −0.34 | −0.01 | 0.13 | 0.01 | 0.19 | −0.05 | −0.28 | 0.15 | −0.34 | −0.04 | 0.68 | | |
| 14. HbA1c annual | 8 | −0.24 | −0.30 | −0.33 | −0.27 | −0.33 | −0.05 | −0.21 | −0.45 | 0.11 | −0.57 | −0.26 | 0.66 | | |
| 15. Lipid profile every 2 years | 9 | −0.32 | −0.32 | −0.48 | −0.39 | −0.37 | −0.11 | −0.15 | −0.38 | 0.24 | −0.55 | −0.28 | 0.63 | | |
| Care area: immunization | | | | | | | | | | | | | | | |
| 16. Influenza (Immunization) | 5 | −0.05 | 0.38 | −0.16 | −0.20 | −0.31 | −0.58 | −0.09 | 0.15 | 0.03 | 0.15 | −0.07 | 0.38 | −0.06[a,b] | 0.36 |
| 17. Pneumococcal (Immunization) | 5 | 0.18 | 0.38 | −0.38 | −0.39 | −0.31 | −0.58 | −0.09 | 0.15 | 0.03 | 0.42 | −0.06 | 0.39 | | |
| Between-measures ANOVA for complexity scales | $F_{(17, 172)}$* | 1.30 | 0.58 | 0.77 | 1.29 | 1.06 | 2.06 | 0.62 | 1.40 | 1.47 | 3.02 | 2.17 | | 3.57 | |
| | *p*-value | 0.20 | 0.90 | 0.72 | 0.21 | 0.40 | 0.01 | 0.87 | 0.14 | 0.11 | 0.0001 | 0.006 | | 0.015 | |

[a]*FJA Functional Job Analysis, EPRP External Peer-Review Program, Avg average, Dx diagnosis, Flex Sig flexible sigmoidoscopy, HTN hypertension, Bp blood pressure, P People, T Things, D Data, WI Worker Instructions, R Reasoning, M Math, L Language, WT Worker Technology, WINT Worker Interaction, HEC Human Error Consequence. Higher numbers indicate greater complexity*

[b]*Within a single column of scores, values with superscripts of different letters (depicted in bold) are significantly different from one another. For example, measures of chronic care post-acute myocardial infarction (superscript a) are significantly higher in measure difficulty than diabetes measures (superscript b); however, immunization measures (superscripts a,b) do not significantly differ from others in measure difficulty. Unbolded items do not statistically differ from each other or from the bolded items*

* *F(3, 186) for the ANOVA by care area*

**Table 3 Pearson Correlations between Number of Tasks in Performance Measures and Functional Job Analysis Scale Ratings ($n$=18 Performance Measures)**

| Complexity scale | Pearson correlation (with n of tasks in measure) | $p$ value |
|---|---|---|
| Complexity scale | | |
| People | −0.210 | 0.404 |
| Things | −0.041 | 0.871 |
| Data | −0.257 | 0.304 |
| Worker instructions | −0.141 | 0.578 |
| Reasoning | −0.244 | 0.330 |
| Math | −0.168 | 0.505 |
| Language | −0.418 | 0.084 |
| Worker technology | −0.487 | 0.040 |
| Worker interaction | −0.254 | 0.309 |
| Human error consequence | −0.114 | 0.653 |
| Composite difficulty score* | −0.297 | 0.232 |

*Spearman's rank order correlation coefficients were comparable (the only significant association was with worker technology, rho=−0.481, p=0.043)*
**Z-scored composite of the ten individual complexity scales*

one-to-one correspondence of a single performance measure with a single difficulty score. The first involved three sets of measures: hypertension control (measures 8 and 9 in Table 2), foot inspections/pedal pulses (measures 10 and 11), and hemoglobin (Hb) A1C control (measures 12 and 13). We chose HbA1C to illustrate the phenomenon in all three sets of measures. The data set included two measures specifying two different cutoffs (HbA1C >11 and HbA1C <9). Nonetheless, the tasks involved in accomplishing the measures were identical, as were their respective task complexity, task difficulty, and measure difficulty scores. In other words, a single score described the difficulty of two different measures. For completeness, however, we reported the two measures independently.

The second instance involved colorectal-cancer-screening measures. We found that performing either full colonoscopy or a flexible sigmoidoscopy could satisfy the measure, which involved two different sets of tasks resulting in two different measure difficulty scores. In other words, one clinical-performance measure was described by two measure difficulty scores. We elected to report both variations of the measure and used both measures in our analyses (see measures 5 and 6 in the Screening section of Table 2). Subsequent results are based on 18, rather than the original 17 measures.

## Descriptive Statistics

Table 2 presents mean standardized individual and aggregated measure complexity and measure difficulty ratings for each clinical-performance measure across clinical care areas. Unstandardized values are provided in Appendix C (available online). Mean ratings for the clinical-performance measures did not exceed medium levels of complexity for any of the individual FJA scales; for the Things and Math dimensions, scores always indicated low levels of complexity. Clinical-performance measures varied substantially in their component number of tasks (mean = 10.56, 95 % $CI_{mean}$ = 7.45 to 13.67; $SD$ = 6.25, 95 % $CI_{SD}$ = 4.69 to 9.37; min = 5, max = 25). The component number of tasks per measure was not significantly related to either the measure difficulty or individual FJA scale complexity ratings (*all ps > 0.05*, with the exception of worker technology, where $r = -0.49$, $p = 0.04$; see Table 3).

## Difficulty Rating Differences among Clinical-Performance Measures and Care Areas

Table 2 presents our analyses of variance. Significant differences existed among measures in measure difficulty, $F_{(17,\ 172)} = 2.17$, $p = 0.006$. Beta blockers and Major Depressive Disorder screening each exhibited higher measure difficulty scores than cervical-cancer screening (pairwise $p = 0.014$ and 0.019, respectively). Additionally, significant differences existed in measure difficulty when grouped by care area, $F_{(3,\ 186)} = 3.57$, $p = 0.015$. On average, measures of chronic care following acute myocardial infarction (i.e., patient receives aspirin and beta blockers at most recent visit) exhibited significantly higher difficulty scores than diabetes or screening measures (pairwise $ps = 0.015$ and 0.001, respectively), though they did not significantly differ from immunization measures.

## Differences in Scale Complexity Ratings among Clinical-Performance Measures

With the exception of math and human-error consequence, no significant differences existed across measures in individual FJA scale complexity scores. Beta blockers at the most recent visit exhibited higher math complexity ratings than either breast-cancer or cervical-cancer screening (pairwise $ps = 0.03$ and 0.01, respectively). Cervical-cancer screening exhibited lower human-error-consequences complexity scores than aspirin at the most recent visit, colorectal-cancer screening (colonoscopy), and Depression screening (pairwise $ps = 0.02$, 0.01, and 0.002, respectively).

# DISCUSSION

We used a novel methodology, adopted from industrial/organizational psychology, for assessing the difficulty of clinical-performance measures as a function of the complexity of their component requisite tasks; this method may help identify the role of individual performance measures in larger strategies for assessing provider and healthcare facility performance, and in fostering learning healthcare organizations. We hypothesized measures would significantly vary from one another in their difficulty scores, thereby demonstrating the method's viability for assessing the work-based difficulty of clinical-performance measures. Our method successfully differentiated between clinical measures of greater versus lesser

difficulty with significant differences in several measures and care areas studied. The method also highlighted the fact that measure difficulty is not a simple matter of work volume. As evidenced in the correlational analysis, the number of tasks required is neither related to the measure's complexity score in any given FJA dimension (for all but one dimension), nor to the measure difficulty score. The data support the assertion that difficult clinical performance measures require great effort, not simply because there are many tasks involved, but because, on average, the tasks' difficulty is high. Thus, the number of steps *and* each step's difficulty must be considered when designing potential interventions for improving quality in a given measure or care area.

To our knowledge, this is the first application of a methodology for characterizing the difficulty of completing clinical-performance measures as a function of the complexity of their component tasks in a primary care context. Describing and assessing the difficulty of the work health-care personnel must do to successfully meet performance measures offers a more nuanced way of understanding the complexity of primary-care performance measures, which in turn could help decision makers choose wisely among the thousands of clinical-performance measures currently available.[30] For example, performance measure difficulty could help identify measures unlikely to yield information about quality of care (e.g., measures so easy everyone can meet them, thus having little to no variance), or help design the length of visits to match the need for clinician time and effort to accomplish the most difficult performance measures. The same approach could be used to more adequately reward areas of clinical care that require more effort, perhaps as a framework for payment of providers replacing the RBRVUs method.

## LIMITATIONS

One limitation is that we examined only a small sample of clinical-performance measures for outpatient primary care; we might have found different results for inpatient medical or surgical measures. Additionally, we only had adequate power to detect very large differences (Cohen's *d*s as large as 1.72) among measures or care areas with a very small number of tasks (e.g., tobacco use $n=5$, influenza $n=5$, pneumococcal $n=5$, immunization care area $n=10$). The fact that there was significant variation across measures in such a small, range-restricted sample suggests the methodology can successfully discriminate among levels of difficulty and is worthy of further examination.

A second limitation is the possibility that the number of tasks in a given measure could vary by site, a possibility for which we could not test. Earlier multi-site work using FJA, by Hysong and colleagues,[11] found no site variation in the set of tasks comprising primary care work, suggesting that site variation is unlikely or quite limited.

## CONCLUSIONS/FUTURE DIRECTIONS

We conclude that difficulty of primary-care measures can be assessed dependably using Functional Job Analysis; the results can provide useful new ways for health-care managers to make decisions about workflow, incentives, and similar administrative concerns that impact the quality of health care.

***Acknowledgements:***

***Contributors:*** *The authors wish to thank Ms. Khai-El Johnson, without whom data collection would not have been possible.*

***Corresponding Author:*** *Sylvia J. Hysong, Ph.D.; Center for Innovations in Quality, Safety, and EffectivenessMichael E. DeBakey VA Medical Center, 2450 Holcombe Blvd. Suite 01Y, Houston, TX 77021, USA (e-mail: hysong@bcm.edu).*

***Compliance with Ethical Standards:***

***Funders:*** *This research was supported by the U.S. Department of Veterans Affairs, Health Services Research and Development Service grants no. CDA 07–0181, PPO 09–278, CIN 13–413 and HFP 90–020.*

***Prior Presentation:*** *The work presented in this manuscript has not been presented elsewhere.*

***Conflict of Interest:*** *Hysong: No conflicts*
*Amspoker: No conflicts*
*Petersen: No conflicts*

## REFERENCES

1. **Chassin MR, Loeb JM, Schmaltz SP, Wachter RM.** Accountability measures–using measurement to promote quality improvement. N Engl J Med. 2010;363(7):683–8. doi:10.1056/NEJMsb1002320.
2. Committee on Quality Measures for the Healthy People Leading Health Indicators. Toward quality measures for population health and the leading health indicators. Washington DC: National Academies Press; 2013.
3. **Hysong SJ, Kelly PA, Woodard LD, Petersen LA.** Prioritizing quality of care in VA medical centers: should we concentrate on diseases or processes? Washington DC: AcademyHealth Annual Research Meeting; 2008.
4. **Brannick MT, Levine EL.** Job analysis: methods, research, and applications for human resource management in the new millenium. Thousand Oaks: Sage Publications; 2002.
5. **Maynard DC, Hakel MD.** The effects of objective and subjective task complexity on performance. Twelfth Annual Conference of the Society for Industrial and Organizational Psychology. Apr 11; St. Louis, MO 1997 p. 1–46.
6. **Wood RE.** Task complexity: definition of the construct. Organ Behav Human Decis Process. 1986;37:60–82.
7. **Schmidt FL, Hunter JE.** The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. Psychol Bull. 1998;124(2):262–74.
8. **Hysong SJ, Best RG, Pugh JA, Moore FI.** Are we under-utilizing the talents of primary care personnel? A job analytic examination. Implement Sci. 2007;2(10):1–13. doi:10.1186/1748-5908-2-10.
9. **Hysong SJ, Amspoker A, Khan MM, Johnson K, Gribble G.** VISN 6 Ambulatory Care System Redesign Improvement Capability Project - Evaluation: Final Report for Fiscal Year 2011. 2011 Sep 21. Report No.: Report to the VA Mid Atlantic Healthcare Network.
10. **Woodard LD, Urech T, Landrum CR, Wang D, Petersen LA.** Impact of comorbidity type on measures of quality for diabetes care. Med Care. 2011;49(6):605–10. doi:10.1097/MLR.0b013e31820f0ed0.
11. **Woodard LD, Landrum CR, Urech TH, Wang D, Virani SS, Petersen LA.** Impact of clinical complexity on the quality of diabetes care. Am J Manag Care. 2012;18(9):508–14.
12. **Hysong SJ, Khan M, Petersen LA.** Passive monitoring versus active assessment of clinical performance: impact on measured quality of care. Med Care. 2011;49(10):883–90.
13. **Fine SA, Cronshaw SF.** Functional job analysis: a foundation for human resources management. Mahwah: Lawrence Erlbaum Associates; 1995.

14. **Asch S, McGlynn E, Hogan M, Hayward R, Shekelle P, Rubenstein L, et al.** Comparison of quality of care for patients in the veterans health administration and patients in a national sample. Ann Intern Med. 2004;141(12):938–45.
15. **Damschroder LJ, Robinson CH, Francis J, Bentley DR, Krein SL, Rosland AM, et al.** Effects of performance measure implementation on clinical manager and provider motivation. J Gen Intern Med. 2014. doi:10.1007/s11606-014-3020-9.
16. **Hofer T, Krein S, Davis J, Hayward R.** Variation at provider, team and facility level for profile measures related to diabetes care. 3rd International Conference on the Scientific Basic of Health Services Research. 2000.
17. **Hsiao WC, Braun P, Yntema D, Becker ER.** Estimating physicians' work for a resource-based relative-value scale. N Engl J Med. 1988;319(13):835–41. doi:10.1056/NEJM198809293191305.
18. **Fine SA.** Functional job analysis. J Pers Adm Ind Relat. 1955;2:1–16.
19. **Fine SA, Getkate M.** Benchmark tasks for job analysis: a guide for functional job analysis (FJA) scales. Mahwah: Lawrence Earlbaum Associates; 1999.
20. **Borman WC.** The occupational information network: an updated dictionary of occupational titles. Mil Psychol. 1996;8(3):263–5.
21. **Peterson NG, Mumford MD, Borman WC, Jeanneret PR, Fleishman EA, eds.** An occupational information system for the 21st century: The development of O*NET. Washington, DC: American Psychological Association; 1999.
22. **Moore F.** Assessment of the human resources for health planning, training and management requirements for achieving the health reforms of law 100. Cambridge: Harvard School of Public Health, Center for Health Economics; 1995.
23. Health Information Technology for Economic and Clinical Health Act, Title XIII of Division A and Title IV of Division B of the American Recovery and Reinvestment Act of 2009 (ARRA). Pub. L. No. 111–5, 123 Stat. 226, §§300jj *et seq*, §§17901 *et seq*, 2-13-2009, 42 U.S.C.
24. Patient Protection and Affordable Care Act. P.L. 111–148 124 Stat. 119, 2010, 42 USC 18001.
25. **Barbour R.** Doing focus groups. London: Sage; 2008.
26. **Cronshaw SF, Best RG, Zugec L, Warner M, Hysong SJ, Pugh JA.** A five-component validation model for functional job analysis as used in job redesign. Ergometrika. 2007;4(1):12–31.
27. Office of Quality and Performance. FY 2007 Technical Manual for the VHA Performance Measurement System including JCAHO Hospital Core Measures. Veterans Health Administration. 2007 July 1. Accessed November 18, 2015 : http://vaww.car.rtp.med.va.gov/filedownload.ashx?fid=3504.
28. VHA Office of Analytics and Business Intelligence. Electronic Technical Manual for the VHA Performance Measurement System. Veterans Health Administration. 2013 November 1. Accessed November 18, 2015: http://vaww.rs.rtp.med.va.gov/ReportServer/Pages/ReportViewer.aspx?%2fPerformance+Reports%2fMeasure+Management%2fMeasureSummary&rs%3aCommand=Render.
29. **Best RG, Pugh JA.** VHA primary care task database [CD-ROM]. San Antonio: South Texas Veterans Health Care System; 2006.
30. National Quality Measures Clearinghouse. U.S. Department of Health & Human Services (HHS) Measure Inventory. Accessed November 18, 2015: http://www.qualitymeasures.ahrq.gov/hhs-measure-inventory/browse.aspx.