THOMSON REUTERS

Analysis of the Focus Shift and Relevance of Grant Applications Received in the First Round of the Provocative Questions Initiative

Summary of Methodology and Results

1.0 Background

The National Cancer Institute's (NCI) Provocative Questions (PQ) initiative seeks to encourage new and innovative approaches to address understudied and particularly challenging areas of cancer research. In early 2011, NCI worked with the cancer research community to determine potential PQ areas through sponsored workshops and online submissions of ideas[1]. In collaboration with NCI's Office of Science Planning and Assessment (OSPA), Thomson Reuters evaluated these potential areas by measuring the volume of research publications and prior NIH grants that address these topics. We will refer to the work completed prior to the RFA release as "Phase 1" and work that later examined the applications received as "Phase 2." The results of the Phase 1 study helped inform the planning and establishment of the 24 PQs that were included in two Requests for Applications (RFAs) released in August 2011[2,3].

In Phase 2 we used an automated text similarity calculation to assign numeric values for the (1) Relevance and (2) Focus Shift of the grant applications that were submitted to the two RFAs. Relevance was calculated by comparing the titles and abstracts of grant applications to the text of the summary statement of the corresponding question in the RFA. In addition to calculating the Relevance of grant applications submitted for the RFA, the Phase 1 work involved the calculation of another set of Relevance values. Phase 1 Relevance calculations reflect the number of grants submitted from 2007-2011, prior to the release of the PQ questions. Because grants identified in Phase 1 were not in response to the PQs initiative, we refer to this measurement as "Coincidental Relevance."

Focus shift was measured by comparing the titles and abstracts of the grant application to prior grants submitted to NIH. Two values of Focus Shift were calculated for each grant application submitted to the RFA. The first was relative to the investigator's own prior work ("By-Self") and the second was relative to NIH grants received from other investigators ("General"). **Figure 1** illustrates the general scheme for calculating Focus Shift and Relevance. More details on the two Focus Shift measurements are presented later in this section.

Table 1 lists the spreadsheets where more detailed information on the specific Relevance and Focus Shift values can be found.[4] The sub-report titled, "Relevance: Most Relevant Applications with Phase1 Merged" required that we repeat the Phase 1 similarity matching to general applications for five PQs that were not matched in Phase 1. Graphic analysis was applied to the text distance data; details are provided in Section 4.0.

[1] NCI Provocative Questions Community Dialog. http://provocativequestions.nci.nih.gov/community-dialog. Accessed April 25, 2011.
[2] RFA: Research Answers to NCI's Provocative Questions (R01). http://grants.nih.gov/grants/guide/rfa-files/RFA-CA-11-011.html. Accessed April 26, 2011.
[3] RFA: Research Answers to NCI's Provocative Questions (R21). http://grants.nih.gov/grants/guide/rfa-files/RFA-CA-11-012.html. Accessed April 26, 2011.
[4] Spreadsheets in **Table 1** were delivered to CSSI and can be found on the SharePoint website for this project.

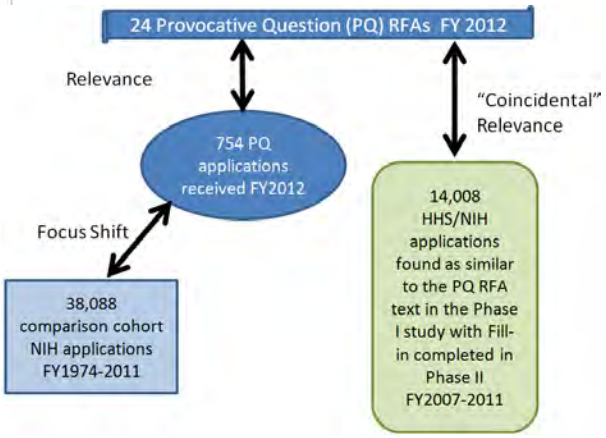

Figure 1. Schematic of Relevance and Focus Shift Calculations.

Table 1. The primary sub-reports resulting from the Thomson Reuters evaluation. The last two columns (Tab 1 Row Count and Tab 2 Row Count) indicate which values of the spreadsheets can be viewed as unique to aid in interpretation.

Excel File Name of Report	Description	Reporting Level (Unique Items)	Tab 1 Row Count	Tab 2 Row Count
Novelty[5]: Summary By Question	A summary of the Focus Shift of applications received for each Provocative Question RFA, with Focus Shift measured against two sets of prior applications - those by the same PI/MPI ("By Self") and a larger "General Set"	Provocative Question (PQ)	Summary (24)	N/A
Novelty: Details by Application	The measure of the Focus Shift for each Provocative Question application against that application's By Self and General Set of prior applications	PQ Application (App)	Details (754)	N/A
Novelty: Most Similar Prior Applications (Focus Shift "Spoilers")	The top four most similar prior applications for each RFA response (By Self and General)	PQ App, Prior App Pair	Most Similar By Self (26846)	Most Similar General (3016 = 754*4)
Relevance: Summary by Question	A summary of the Relevance of applications received for each Provocative Question RFA	Question	Summary (24)	N/A
Relevance: Details by Application	The measure of the Relevance of each Provocative Question application	PQ Application	Details (754)	N/A
Relevance: Most Relevant Applications with Phase 1 Merged	The top most relevant RFA applications merged with Phase 1 similar applications. All RFA applications are included with their rank in the merged list. No more than 100 of the most similar Phase 1 applications are shown.	Question, PQ App or Phase 1 App Pair	Most Relevant (2621)	N/A

All measures were based on text similarity as calculated by a standard algorithm: the version of the Okapi BM25 query-to-document similarity score underlying the Microsoft SQL Server™ Full Text Search feature. Measurements shown are generally expressed as text distances in the range from 0 (similar) to 1 (dissimilar). The text distance scale from 0 to 1 is illustrated in **Figure 2**, along with a suggested interpretation text distance as a measure of Relevance and Focus Shift. In some instances it is helpful to

[5] In early phases of the study, the phrase "focus shift" was called "novelty", and the earlier terminology remains in some areas of the Excel reports as of January 2013.

[6] In some cases, less than 4 prior NIH applications were found to be associated with the PI or MPIs that submitted the PQ Application. In 52 cases, no prior NIH applications were found.

have both Relevance and Focus Shift increasing with the coordinate scale, as in the case of scatterplots graphing Relevance against Focus Shift. In these cases, one of the measurements is converted to a similarity measure using the simplest linear conversion: similarity = 1 – distance.

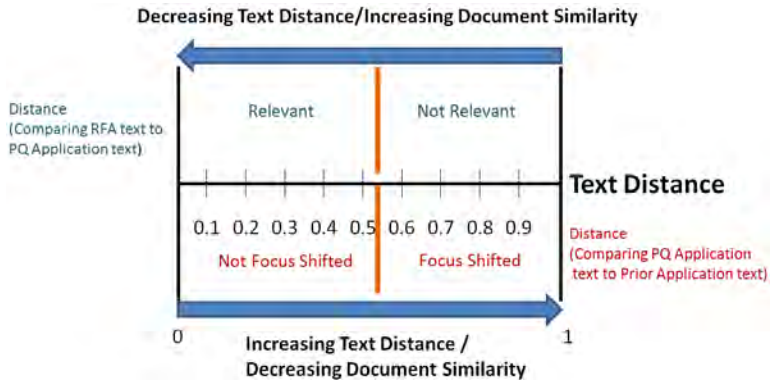

Figure 2. Schematic illustrating the correlation between text distance and our current interpretation of its relationship to grant application Relevance and Focus Shift. The orange line denotes the current threshold for both.

An important technical point is that the text similarity measurements are based on a Term Frequency/Inverse Document Frequency calculation (TF/IDF). As such, the values that result from the calculations are dependent upon the selection of relevant documents that provide a corpus of text. The two sets of Relevance measurements were made using a single document corpus of the PQ applications merged with the Phase 1 similar applications so the Relevance measures between the two classes of applications could be directly compared. The Focus Shift measurement was made using the single document corpus formed by the comparison cohort[7], but for each PQ application, the measurements were partitioned into two subsets based on whether the prior comparison application was submitted by the same investigator (defined as the "By-Self" subset) or by different investigators (defined as the "General" subset). The diagram in **Figure 3** illustrates this partition for a single PQ application.

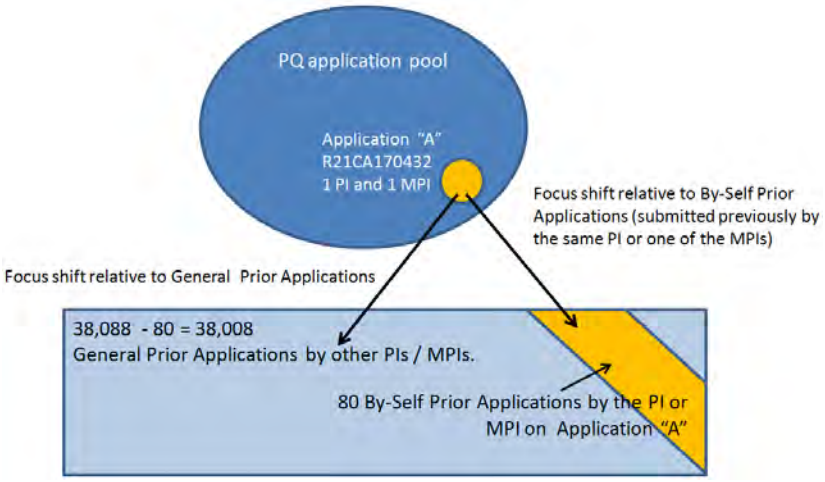

Figure 3. Diagram of Partition of the Document Corpus for Focus Shift Calculations.

[7] Supplemented by the text of the 754 PQ applications to allow the self-similarity scoring used for Focus Shift General distance scaling.

2.0 Methodology

2.1 Defining an appropriate comparison cohort for Focus Shift and ranking of PQ RFA applications to prior applications submitted by PQ applicant, or to those submitted to NIH broadly.

2.1.1 General Rules

1. The starting set for prior application selection includes all applications in IMPAC II as of December 2011.
2. Only prior applications to NIH were used to measure Focus Shift of PQ applications; grant applications to other DHHS agencies were excluded.
3. The comparison cohort of prior applications used to measure Focus Shift consists of a single pool of applications. However, for each PQ application, the association to this pool was dynamically re-partitioned into two subsets: the By-Self applications found by searching for prior applications submitted by the same PI (or MPIs if configured – see below), and a General subset that includes all applications submitted by any other PI in the search sample. For a given PQ application, both the By-Self and General sets have additional restrictions that are described in detail below. Currently, the restrictions are the same for both subsets, except for the set of individuals allowed as PI/MPI on a prior application.
4. Note that:
    A. For a given PQ application, the By-Self and General subsets do not overlap – a prior application is either one or the other.
    B. For a given PQ application, the corresponding General prior applications can overlap with the By-Self prior applications for a different PQ application.
    C. The By-Self subsets for2 PQ applications - “A” and “B” - can overlap, since there are 42 individuals that appear on more than one PQ application (measured using both PIs and MPIs on all applications).

2.1.2 Table-driven Configuration Items:

1. Single parameters

Table 2. Description and values of single parameters

Short Description	Long Description	Current Value (4/9/12)
NCI PI Sampling Pct	Percent of individuals in IMPACII who have submitted either NCI applications only or who have submitted most of their applications to NCI, and who are selected to form the set of General prior applications	50%
Non-NCI Sampling Pct	Percent of all other individuals in IMPACII (submitted no NCI applications or submitted most applications to other ICs) who are selected to form the set of General prior applications	6%
Max FY	Last Fiscal Year for inclusion of prior applications in the comparison cohort	2011
Min FY	Earliest Fiscal Year for inclusion of prior applications in the comparison cohort	2007
Use PQ MPIs?	For each PQ application, only the contact PI name is used to search for prior applications. If this value is “Yes,” this indicates that the MPI name(s) on the PQ application should also be included in the search.	Yes
Minimum Abstract Length	The minimum number of characters required for an abstract to be recognized as associated with an application	500
Focus Shift Threshold for By-Self	The distance from a PQ application to all of its By-Self prior applications must exceed this value for the PQ application to be rated as Focus Shifted relative to the By-Self subset	0.53
Focus Shift Threshold General	The distance from a PQ application to all of its General prior applications must exceed this value for the PQ application to be rated as Focus Shifted relative to the General subset	0.53
Relevance Threshold	The distance from a PQ application to the RFA test must be less than or equal to this value for the PQ application to be considered Relevant	0.53
Use Self Ranks for Relevance 0 Distance	Compare PQ application text to itself to establish the rank that is translated to 0 distance when measuring Relevance	No
Focus Shift By Self Zero Distance	Indicates which of several possible maximum similarity scores to convert to a zero value for Focus Shift by self text distance. "PerLeftMaxRank" means: Use the largest score obtained for each PQ App.	PerLeft MaxRank
Focus Shift General Zero Distance	Indicates which of several possible maximum similarity scores to convert to a zero value for Focus Shift by self text distance. "SelfRank" means: Use the largest score obtained for each PQ App, including scores obtained by comparing the PQ app text to itself.	SelfRank
Search String Length	Maximum number of characters used on the left side of a text distance measurement (required for performance).	4000
Top N for Most Similar Prior report	Number of the most similar (Focus Shift Spoiler) prior applications to show for each PQ application in that report	4
Top N for Most Relevant Application report	Number of the most Relevant PQ applications to show for each Question in that report	100

2. Multi-value parameters
    A. Allowed Activity Codes (for the general prior applications, all codes allowed for the By-Self set)
        i. DP1, DP2, DP3, DP4, DP5, P01, P20, P42, P50, PN1, R00, R01, R03, R15, R21, R22, R23, R29, R33, R34, R35, R36, R37, R41, R42, R43, R44, R55, R56, RC1, RC2, RC3, RC4, RL1, RL2, RL5, RL9, U01, U19, U34, U43, U44, UA5, UC1, UC2, UC4, UC7, UH2, UH3, UM1
        ii. The above codes are currently allowed for both the General and By-Self subsets.
    B. Allowed ICs (for the general prior applications, all codes allowed for the By-Self set)
        i. NIAAA, NIA, NIAID, NIADD, NIAMS, NCCAM, ODNCI, CLC, CIT, NIDA, NIDCD, NIDCR, NIDDK, DS, NIBIB, WT, NEI, GIFT, NIGMS, NICHD, NHGRI, ODHLI, JOINT, NLM, NCMHD, MF, NIMH, NINR, NINDS, ORF, ORS, CSR, RM, NCRR, DRS, FIC, WH
        ii. The above codes are currently allowed for both the General and By-Self subsets.

2.1.3 Restrictions for the By-Self Subset (relative to a given PQ application; rules are applied in the order shown)[8]

1. Determine the set of Individuals for whom prior applications will be found. If the "Use PQ MPIs" parameter is "No", this will be the single contact PI. If the "Use PQ MPIs" parameter is "Yes", this will include both the contact PI and all MPIs (if applicable).
2. Determine the set of NIH applications for the individuals from step (a) for which the individual was either the PI or MPI, always excluding the PQ applications themselves.
3. From the set of applications found in step (b), consider only those applications that satisfy **both** of the following conditions:
    A. Fiscal Year is known and falls between the Min FY and Max FY values (see 2(a), above).
    B. The application must be associated with an entry in the IMPAC II Abstracts table which has an Abstract text length in characters greater than the value of the minimum abstract length parameter (see 2(a) above).

2.1.4 Restrictions for the General Subset

Applications found with the following restrictions are added to the cohort pool (unless they are already present), and allowed to be part of the General subset for any PQ application. The applications already found in Step 2 are allowed or disallowed in the General set, depending on the PQ application being considered.

1. Locate the set of Individuals in IMPACII who have all of their applications (considered as having no restrictions) associated with NCI or who have both NCI and non-NCI applications, but more NCI than non-NCI. Take a 50% sample of this set.
2. Take a 6% sample of everyone else in IMPACII who appears as a PI or MPI on at least one application. Combine with the sample from step 1.
3. Find all applications (not already found) from these sampled Individuals that also meet all of the restrictions (which are more stringent than those used to select the By-Self set):

[8] In some cases, the order of application of a subset of these rules does not change the final result set.

i. Fiscal year is known and is between the Min FY and Max FY values (see 2(a), above).
ii. The application has a known type, which is either Type 1 (new), 2 (competing continuation), or 5 (non-competing continuation).
iii. The application suffix code is either NULL (treated as Amendment 0), or if it is filled in, does not contain "S".
iv. The values for Application Status Code, Priority Score, Activity Code, Serial Number, and IC must all be known (i.e., not NULL), and the Serial Number must be positive.
v. The Activity Code (aka "Mechanism") for the application must be one of those listed in 2(b)i(1) in Section 2.1.2 above.
vi. The application must be associated with an entry in the IMPAC II Abstracts table which has an Abstract text length in characters greater than the value of the minimum abstract length parameter (see 2(a) in Section 2.1.2 above).
vii. Sub-selection Rule (method of choosing a single application to "represent" a related group of applications that could vary by application type, amendment number, or some other unmeasured attribute(s)):
    1. Place all prior applications that have passed the previous selection rules into sub-groups having the same values for IC, Activity Code, and Serial Number (aka "the same triplet").
    2. In each triplet, list the applications in order of the length of their abstracts in characters, longest first, and then break ties by sub-ordering based on the specific FY of each application, most recent first, and finally break ties using the IMPAC II Application ID, highest first.

i. Choose only the application listed first in the ordering from (2) for each triplet.

4. Applications passing all of these restrictions are recorded, assigned the "General" relationship to PQ applications that do not share any of the PIs (or MPIs, if this parameter is included), and their metadata; specifically, the project title and abstract text are collected.

2.2 Data Preparation

1. PQ Application cohort identification
    - The set of 754 applications responding to the PQ RFAs was provided by NCI.

2. Phase 1 data and PQ entity definition (details on specific mapping of Phase 1 to Phase 2 work)
    - Phase 2 work included the calculation of Relevance scores for grant applications to the PQs RFAs. As a comparison group, we re-ran Phase 1 Relevance calculations for all questions that ultimately made it to the PQ RFA. In some cases, questions were merged or new questions were added that were never evaluated in our Phase 1 Study. It was therefore necessary to map the Phase 1 study to the questions that were ultimately presented in the RFA. A map of Phase 1 questions to Phase 2 was provided by NCI, see **Table 3** below.
    - Phase 1 target text and similar applications[9]:

[9] There were several iterations of the Phase 1 work based on OSPA feedback and refinement of the methodology. All iterations were maintained in the Thomson Reuters databases. Of the iterations used in Phase 1, we selected the version with the latest date, or, in the case of ties, the highest number label.

In addition to the summary statement of the proposed question, Relevance scores in Phase 1 work also used abstracts handpicked by OSPA staff as the "Target Text" for ranking prior grant applications. All target text from Phase 1 was added to the aggregate text of the current question to which the Phase 1 question is mapped in **Table 3** (as noted above, not all questions could be mapped between Phases). Additional text was provided by NCI for current question numbers 12 and 21. We will refer to this update as the "Phase 1 Fill-In." The similar applications were re-mapped to the current question numbers in the same way. The application set used in this study was then filtered using the same rules for defining the By-Self and General prior applications, with the following exceptions:

- The Activity Code restrictions were not applied.
- The IC restrictions were not applied (non-NIH HHS applications were allowed).
- There was no PI sampling – all PIs in IMPACII were allowed.

- Phase 1 fill-in for PQs 3, 12, 20, 23, and 24: After remapping from the Phase 1 question list, no similar applications were available for these five PQs, as the searches had not been completed at the end of Phase 1. Thus, search terms were prepared for these five questions by NCI and Thomson Reuters subject matter experts and the text matching methodology from Phase 1 was applied to find similar past applications. The applications were filtered, as explained above, for the other Phase 1 similar applications.

Table 3. Phase 1 Fill-In Mapping

Current Number	Question statement	Phase 1 Number	Short description
1	How does obesity contribute to cancer risk?	5	Obesity and Cancer
2	What environmental factors change the risk of various cancers when people move from one geographic region to another?	11	Environmental Risks Moving Geographic Location
3	Are there ways to objectively ascertain exposure to cancer risk using modern measurement technologies?	-	-
4	Why don't more people alter behaviors that are known to increase the risk of cancers?	12	Altering Behaviors Known to Increase Cancer Risk
5	Given the evidence that some drugs commonly and chronically used for other indications, such as an anti-inflammatory drug, can protect against cancer incidence and mortality, can we determine the mechanism by which any of these drugs work?	10	Off-Label Drugs Prevent Cancer
6	What are the molecular and cellular mechanisms by which patients with certain chronic diseases have increased or decreased risks for developing cancer, and can these connections be exploited to develop novel preventive or therapeutic strategies?	20	Disease Cancer Correlation
7	How does the life span of an organism affect the molecular mechanisms of cancer development, and can we use our deepening knowledge of aging to enhance prevention or treatment of cancer?	19	Age Dependence
8	Why do certain mutational events promote cancer phenotypes in some tissues and not in others?	14	Tumor Development
9	As genomic sequencing methods continue to identify large numbers of novel cancer mutations, how can we identify the mutations in a given tumor that are most critical to the maintenance of its oncogenic phenotype?	16	Driver and Passenger Mutations_20110725
10	As we improve methods to identify epigenetic changes that occur during tumor development, can we develop approaches to discriminate between "driver" and "passenger" epigenetic events?	16	Driver and Passenger Mutations_20110725

11	How do changes in RNA processing contribute to tumor development?	15	Alternative Splicing
12	Given the recent discovery of the link between a polyomavirus and Merkel cell cancer, what other cancers are caused by novel infectious agents and what are the mechanisms of tumor induction?	-	-
13	Can tumors be detected when they are two to three orders of magnitude smaller than those currently detected with in vivo imaging modalities?	8	Tumor Detection Smaller In Vivo
14	Are there definable properties of a non-malignant lesion that predict the likelihood of progression to invasive or metastatic disease?	3	Nonmalignant Tumors to Invasive Cancer
15	Why do second, independent cancers occur at higher rates in patients who have survived a primary cancer than in a cancer-naïve population?	9	Second Primary Cancers
16	How do we determine the clinical significance of finding cells from a primary tumor at another site?	6	Clinical Significance Metastatic Tumors
17	Since current methods to assess potential cancer treatments are cumbersome, expensive, and often inaccurate, can we develop other methods to rapidly test interventions for cancer treatment or prevention?	13	Testing Combination Therapies
18	Are there new technologies to inhibit traditionally "undruggable" target molecules, such as transcription factors, that are required for the oncogenic phenotype?	2	Undruggable Targets
19	Why are some disseminated cancers cured by chemotherapy alone?	1	Cancers Cured by Chemo Only
20	Given the recent successes in cancer immunotherapy, can biomarkers or signatures be identified that can serve as predictors or surrogates of therapeutic efficacy?	?	Description?
21	Given the appearance of resistance in response to cell killing therapies, can we extend survival by using approaches that keep tumors static?	7	Overcoming Tumor Resistance to Radiotherapy
22	Why do many cancer cells die when suddenly deprived of a protein encoded by an oncogene?	4	Oncogene Addiction
23	Can we determine why some tumors evolve to aggressive malignancy after years of indolence?	-	-
24	Given the difficulty of studying metastasis, can we develop new approaches, such as engineered tissue grafts, to investigate the biology of tumor spread?	-	-
-	-	17	Epstein-Barr Virus
-	-	18	Viruses, Bacteria, and Cancer
-	-	21	Gender and Cancer Rates

2.3 Text Distance Measurements

1. For all classes of applications (PQ applications, By Self prior, General prior, Phase 1 similar, and Phase 1 fill-in), the Title and Abstract were combined to form the search text for Relevance and Focus Shift distance measurements, following some text clean-up and standardization. All text distances were measured using a software statement of the form:

$$get\_similarity\_core\ (left\ side\ text, right\ side\ text).$$

The underlying Microsoft® software is not symmetric in how it treats the left and right side text strings, and additional text clean-up and limitation to a fixed length had to be applied to the left side text for program performance (see the parameter list above). The left/right sides used were:

- Focus Shift
  - Left = PQ Application Title+Abstract
  - Right =Comparison Cohort Prior Application Title + Abstract
- Relevance

- Left = Question Statement + Background+Feasibility+Success Implications+Target Text

  (all but the last from the NCI website)
- Right = PQ or Phase 1 Application Title+Abstract

2. **Distance scaling.** The direct measurement of the text produces a similarity score for each pair of documents (Score(Left, Right)). The score has a documented absolute range of 0 to 1000, with the highest score corresponding to the most similar documents. Using the formula below, scores were converted to distance, in which a case with the maximum similarity score is a distance of 0, and a 0 similarity score is the maximum distance of 1.

$$Distance\ (Left, Right) = \frac{\max(all\ Scores\ from\ Left\ \text{to}\ any\ right\ side\ document) - Score(Left, Right)}{\max(as\ above)}$$

Summarized as

Distance = (Max- score)/ Max = (1 – score/Max)

3. **Scaling options.** There are several options for defining the maximum score used in the scaling formula that amount to providing a definition of "any right side document". Three options are illustrated in **Figure 4** as M_1, M_2, and M_3. M_1 defines "any right side document" as any document actually in the right side corpus, even if the left side document is not part of the corpus[10]. M_2 requires that the left side document also be present in the right side corpus to allow a score of the similarity of a document to itself. M_3 extends "any right side document" to include documents from some arbitrarily larger set. For this study, based on QA performed by NCI, the M_2 option (self-scoring) was used for scaling Focus Shift General distances and the M_1 option (non-self max) was used for Focus Shift By-Self distances and Relevance distances.

[10] In this particular study, the left side was never automatically part of the right corpus: PQ applications were new FY 2012 applications and thus not part of the 2007-2011 comparison cohort. To implement the self-scoring for Focus Shift General, the PQ applications were temporarily inserted into the comparison cohort and then removed after calculating the self-scores.

Similarity Score (s) to Text "Distance" (d) scaling

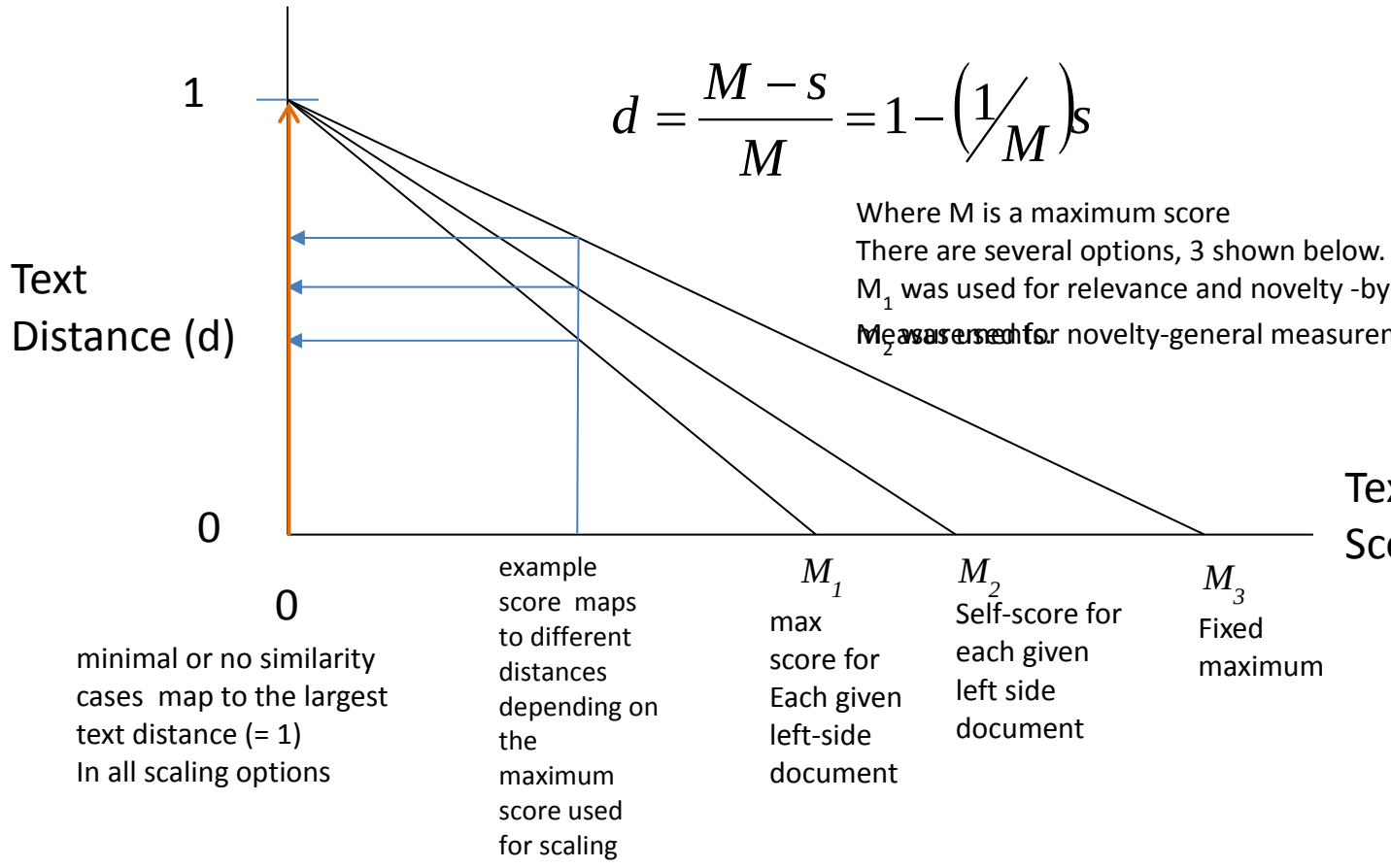

Figure 4. Schematic of the influence of scaling on text distance measurements.

4. **Distance counts.** The table below summarizes the number of documents for which pair-wise similarity scores were measured and scaled into distance values:

Table 4. Summary of document counts in pair-wise similarity measurements

Measure	Left Side Documents	Left Side Count	Right Side Documents	Right Side Count	Relevant Pairs	Irrelevant Pairs	Total Pairs	Comment
Focus Shift	PQ Applications	754	Comparison Cohort Prior applications PQ Applications inserted to allow self-scoring	38,088+754 = 38,842	29,286,868	0	29,286,868	Everything in the comparison cohort is either a General or By Self prior to each PQ app and all self-scores are used

Relevance	RFA Question Text	24	PQ Applications Phase 1 similar applications Fill-in similar applications for PQs missing or incomplete in Phase 1	754+14,008 =14,762	23,298 = 754 RFA, PQ App pairs + +21,644 RFA, Phase 1 app & Fill-in pairs	330,990	354,288	Only interested in comparing an application to its corresponding RFA The irrelevant distances come "for free" and were used for exploratory analysis

2.3 Measurement Quality Assessment (QA) and Focus Shift / Relevance Threshold Selection

Relevance Threshold. Although the Relevance text distance (from the RFA text to a PQ Application or Phase 1 similar application) is expected to be inversely well correlated with an expert's subjective assessment of scientific Relevance, the exact form of the relationship is uncertain. To alleviate this uncertainty, we set the modest goal of being able to determine a single, constant "threshold distance" value for Relevance. Applications with text distances less than the threshold (closest or most similar to the RFA text) would be classified as "Relevant" and all others would be "not relevant", with allowance for some error band around the threshold that could be ignored for summary reporting purposes.

Focus Shift Threshold. PQ Applications were compared to all prior applications by the same investigators (the By-Self subset) and a large number (~38 K) of other prior applications. We defined Focus Shift as an "all-or-nothing" classification, for which high similarity (low distance) to even one prior application in those subsets would disqualify a PQ application from being rated as "Focus Shifted". We set the goal of determining two fixed Focus Shift thresholds – one relative to the By-Self subset and one relative to the General subset. To be classified as Focus Shifted relative to the By-Self subset, a PQ application had to have its "closest approach", or minimum text distance to *any* By-Self prior application above the Focus Shift By-Self threshold value. Similarly, to be classified as Focus Shifted relative to the General subset, the minimum text distance to any General prior application must have exceeded the Focus Shift General threshold.

Threshold QA. To converge upon the best threshold values, we started with a preliminary set of trial thresholds and generated three samples each (Focus Shift General, Focus Shift By-Self, and Relevance) for document pairs in four distinct text distance bands – one near 0 (most relevant, least Focus Shifted), two on either side of the threshold, and one for large distances between the threshold and the maximum distance of 1 (least relevant, most Focus Shifted). The samples included the text used for each of the two documents in each pair, and the scaled distance value between each pair. The samples were provided to NCI, which returned expert assessments of the actual level of Relevance or Focus Shift for some of the pairs in each sample.

These results were used to select the current threshold value – a single text distance value of 0.53 was found to work well as the threshold for Relevance and both Focus Shift classifications. Applications for

which the lowest text distance (“closest approach”) to any prior By-Self or General application is 0.53 or larger have been classified as Focus Shifted relative to the respective subsets. Applications whose distance to the RFA text is 0.53 or less have been classified as relevant to that RFA.

3.0 Analysis Results

In this section we present the results obtained from the text distance measurements, focusing on the graphic analysis. The Excel files listed in Section 2.0 are discussed at a high level; detailed descriptions of the data in these files can be found in the accompanying Variable_List.xlsx.

3.1 Focus Shift

3.1.1 PQ Application Focus Shift

Since it is expected that investigators will tend to carry over ideas from prior research, achieving a Focus Shift classification in comparison to one's previous applications – the By-Self subset – was expected to pose a challenge. Similarly, since the general comparison cohort was constructed with no restrictions on the topics of research, it was expected that a "close approach" by a general prior application would be less likely than for the By-Self subset.

Of the 754 PQ Applications, 39 (5.2%) were classified as Focus Shifted relative to the By-Self prior subset and 271 (35.9 %) were classified as Focus Shifted relative to the General prior subset. The numbers and percentages of Focus Shifted applications in response to each RFA are summarized in **Table 5**.

Table 5. Classification of PQ Applications by Focus Shift Rule

Question	Applications Received	Focus Shifted Relative to By-Self	Focus Shifted Relative to General
1	84	1 (1.2%)	33 (39.3%)
2	15	1 (6.7%)	8 (53.3%)
3	12	0 (0%)	11 (91.7%)
4	15	0 (0%)	10 (66.7%)
5	67	2 (3%)	27 (40.3%)
6	31	1 (3.2%)	12 (38.7%)
7	19	0 (0%)	7 (36.8%)
8	19	1 (5.3%)	3 (15.8%)
9	31	1 (3.2%)	14 (45.2%)
10	27	1 (3.7%)	10 (37%)
11	50	2 (4%)	16 (32%)
12	28	1 (3.6%)	16 (57.1%)
13	22	1 (4.5%)	14 (63.6%)
14	50	4 (8%)	22 (44%)
15	8	2 (25%)	3 (37.5%)
16	9	0 (0%)	1 (11.1%)
17	32	6 (18.8%)	10 (31.3%)
18	69	4 (5.8%)	25 (36.2%)
19	9	0 (0%)	2 (22.2%)
20	31	0 (0%)	8 (25.8%)
21	42	3 (7.1%)	9 (21.4%)
22	24	1 (4.2%)	2 (8.3%)
23	23	4 (17.4%)	5 (21.7%)
24	37	3 (8.1%)	3 (8.1%)

The full set of distributions of minimum distances (closest approaches) of PQ applications for each question to their set of By-Self prior applications is shown in **Figure 5**. The portion of the distribution to the ***right*** of 0.53 represents the applications that were Focus Shifted for each question (an analogous graph exists for Relevance in which relevant PQ applications will be represented by the distribution to the ***left***).

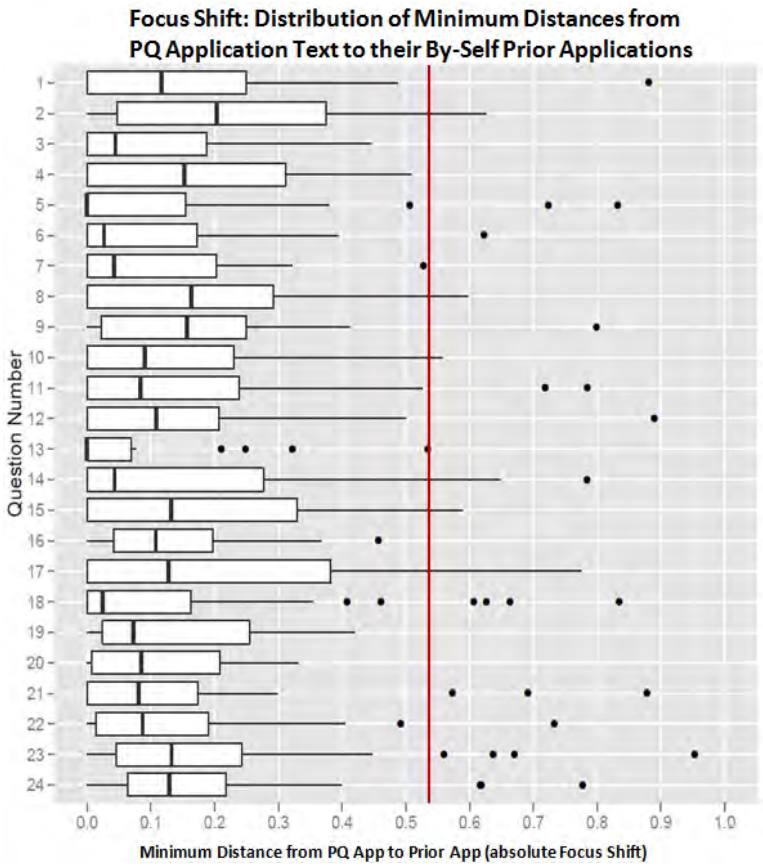

Figure 5. Distributions of PQ Application Focus Shift Versus Prior Applications of Applicant, by PQ.

The full set of distributions of minimum distances (closest approaches) of PQ applications for each question to their set of general prior applications is shown in **Figure 6**. The portion of the distribution to the ***right*** of 0.53 represents those applications that were Focus Shifted relative to the general prior applications for each question. Most PQ applications achieved Focus Shift in this sense, with few exceptions (26). These exceptions can be further investigated by using the **Novelty_Details_by_Application.xlsx**[11] file and applying a filter on "Is Focus shifted compared with general prior applications?" (Column N) = No to view the applications that were not Focus Shifted relative to the general prior applications.

[11] In early phases of the study, focus shift was called "novelty"; that terminology remains in some areas of the Excel report files as of January 2013

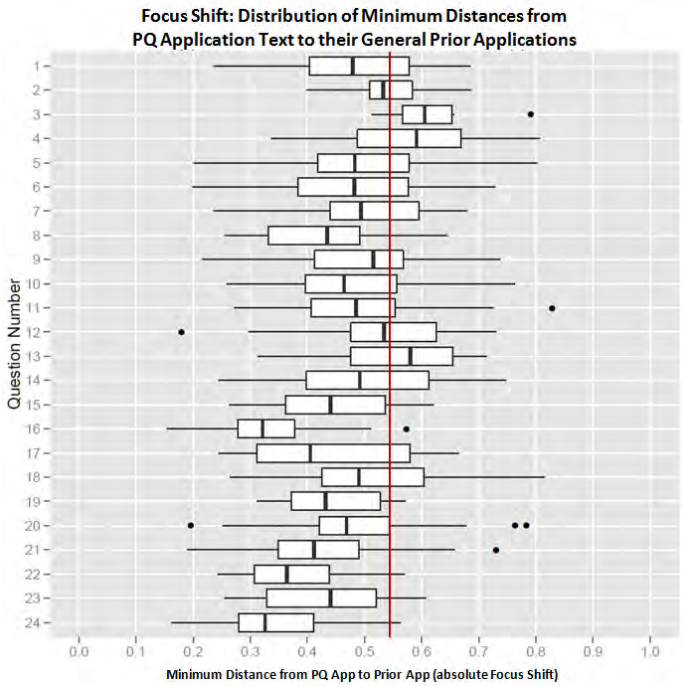

Figure 6. Distributions of PQ Application Focus Shift Versus Prior NIH Applications, by PQ.

3.1.2 Subdividing PQ Applications by Extent of Focus Shift

To better understand the correlations between Focus Shift, as measured by the automated text similarity algorithm, and similarity of science between two grant applications, we conducted a manual review using subject matter experts. This effort was carried out using a subset of 40 grant applications with very low similarity scores; distance <0.05. We refer to these as possible "repurposed applications." We first examined the full set of applications that met the criteria, then assigned the applications to two subgroups based on the nature of the prior application to which they were most similar. This resulted in the following three groups for analysis:

1. All applications
2. Applications repurposed from unfunded prior grants
3. Applications repurposed from funded prior grants that resulted in publications

Manual review of the 40 repurposed applications showed two general trends. First, results for grants repurposed from unfunded applications indicated that:

- 55% had reused prior application text
- 30% reused background text, but the experimental approach was found to be significantly different
- 15% were extensions to prior applications

Second, results for grants repurposed from unfunded applications indicated that:

- 25% had reused prior application text
- 30% reused background text, but the experimental approach was found to be significantly different
- 45% were extensions to prior applications

Our findings in the manual review suggest that PQ applications with low similarity scores in the current implementation of our text similarity algorithm cannot be assumed to have been reused from prior grants.

Using the threshold of Relevance Scores < 0.05, we found that 41% (311/754) of applications to the PQs were potentially repurposed; over 50% of applications to PQs 3,5,13,14 and 15 met this criteria. Tables in Appendix 1 show the percentages of potentially repurposed PQ grant applications for each of the four categories. For the remaining two subcategories, 25% (189/754) grants were repurposed from unfunded prior grants, and 12% (88/754) were repurposed from funded prior grants with publications.

3.1.3 Focus Shift Data Files

To examine in more detail the closest approach to each PQ application by By-Self or General prior applications, refer to the **_Most_Similar_Prior_Applications.xslx** file. This file shows the four By-Self and four General prior applications with the lowest text distance for each PQ application, along with the title and abstract text of the PQ application and the prior application. For cases of *non*-Focus Shifted applications (relative to By-Self or General), at least one of the 4 distances will be less than the 0.53 threshold, and those prior applications represent the "Focus Shift spoilers" that resulted in it failing the test for a minimum text distance above the threshold. For Focus Shifted PQ applications, the top four prior application distances give a measure of the "margin" for the closest approach.

The QA files also provide the text of the prior applications that were used for comparison to a given PQ application's text for selected cases in the QA distance bands (i.e., sample text from prior applications was provided to represent small, medium and large distances relative to each PQ grant application). For other specific cases of a PQ application to prior application comparison (below the top four most similar and outside the QA samples), Thomson Reuters can provide a detailed report on request.

The report package also contains graphs (summary boxplots and graphs of an estimated probability density function) of all distances, rather than just the minimum distance to the prior application subsets. The question-level summary distribution appears in the PDF files whose names start with "Novelty_Distances". The distribution for each PQ application can be viewed by following a link in the **Novelty_Details_by_Application.xlsx** file. When there are only a small number of By-Self prior applications, the distribution is shown as a simple bar chart. There is no data for the By-Self Focus Shift measurement for 139 PQ applications, since no prior applications were found by the same PI or MPIs. In those cases, the entire comparison cohort was used for the general Focus Shift measurement.

3.2 Relevance

3.2.1 PQ Application Relevance

Of the 754 PQ Applications, 614 (81.4%) were classified as scientifically relevant to the topic established in the RFA. The numbers and percentages of relevant applications in response to each RFA are summarized in **Table 6**.

Table 6. Relevance of PQ Applications to RFA Text, by PQ Number.

Question	Applications Received	Relevant
1	84	75 (89.3%)
2	15	14 (93.3%)
3	12	11 (91.7%)
4	15	10 (66.7%)
5	67	49 (73.1%)
6	31	25 (80.6%)
7	19	17 (89.5%)
8	19	19 (100%)
9	31	26 (83.9%)
10	27	20 (74.1%)
11	50	41 (82%)
12	28	19 (67.9%)
13	22	14 (63.6%)
14	50	41 (82%)
15	8	5 (62.5%)
16	9	9 (100%)
17	32	27 (84.4%)
18	69	57 (82.6%)
19	9	9 (100%)
20	31	12 (38.7%)
21	42	39 (92.9%)
22	24	23 (95.8%)
23	23	19 (82.6%)
24	37	33 (89.2%)

The full set of distance distributions of PQ applications to the topic text of their corresponding question's RFA is shown in **Figure 7**. The portion of the distribution to the ***left*** of 0.53 represents the applications that were relevant to the RFA text for each question. The graph shows a high degree of variability in Relevance among the applications for particular questions and significantly different distributions across the PQs. One possible conclusion is that Relevance is more difficult to measure than Focus Shift using text comparisons, suggesting that it would be helpful to consider the comparative measures of the Relevance of PQ applications and the "coincidental" Relevance of the set of Phase 1 similar applications, which is discussed in Section 4.2.2.

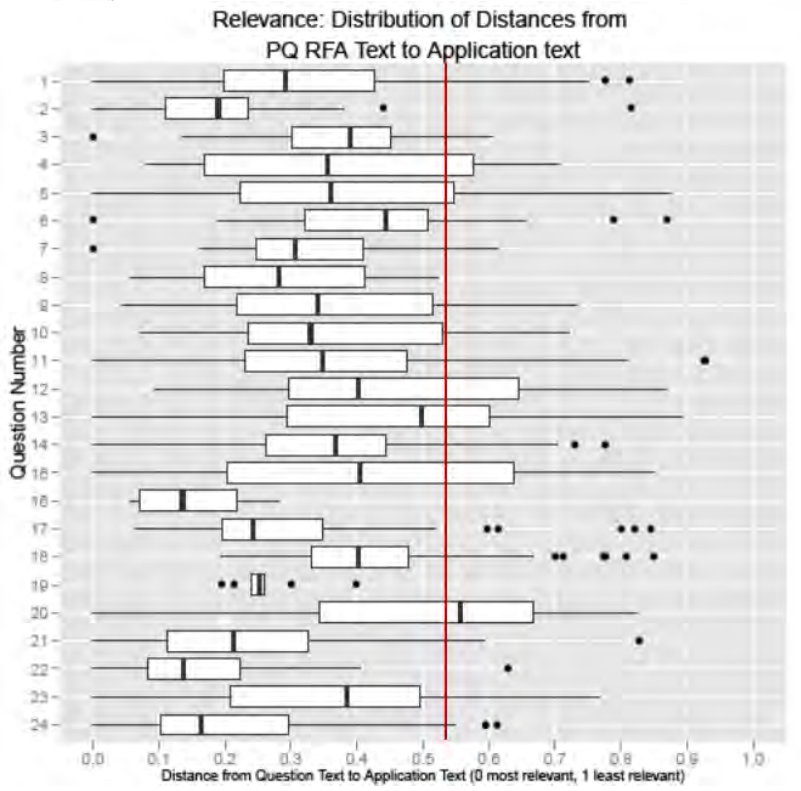

Figure 7. Relevance of PQ Application Text to RFA Text, by PQ.

All of the 754 distance measurements for each PQ application represented in the boxplots above are available in the **Relevance_Details_by_Application.xslx** file, which also has indicators for whether an application passed the Relevance threshold and/or was advanced to the Face-to-Face review. The application text and question text is also available in this file for all measurements.

3.2.2 PQ Applications Compared with Phase 1 Similar Applications

In this section, we compare the measured Relevance of the PQ applications to the Coincidental Relevance of one of the applications found in Phase 1 by searching past grant applications for sets of terms that were customized to yield good match results for each PQ in the RFA.

In this comparison, it should be noted that, unlike the PQ application set, in which each PQ was assigned to a single PQ (by PI assertion or NCI determination), the set of Phase 1 similar applications[12] were matched to all PQs for which a match was possible according to the parameters of the search. **Table 7** shows the cardinality of the matching of Phase 1 similar applications to the Phase 1 PQ questions.

[12] From this point on, any reference to "Phase 1 similar applications" should be understood to include the fill-in applications matched during Phase 1, with Phase 1 similar applications appropriately re-mapped to the new set of question numbers.

Table 7. Matching of PQ Applications to PQ RFA Text.

Number of questions matched to a given application	Number of Phase 1 similar applications with this number of matched questions	Question/Application pairs
1	8,612	8,612
2	3,735	7,470
3	1,219	3,657
4	334	1,336
5	84	420
6	20	120
7	3	21
8	1	8
	14,008	**21,644**
	Total Distinct Phase 1 applications	Total pairwise Question/Application assignments

When measured for Relevance alongside the PQ applications (using the same corpus and scaling rules), 11,280 of the 21,644 Phase 1 similar applications were coincidentally relevant (52.1%). The number and percentage of the relevant Phase 1 similar applications for each question are shown in **Table 8**.

Table 8. Number of PQ Applications Relevant to Phase 1 PQ Text, By Question.

Question	Phase 1 Similar Applications	Relevant
1	464	301 (64.9%)
2	129	81 (62.8%)
3	2015	318 (15.8%)
4	885	469 (53%)
5	55	45 (81.8%)
6	1206	585 (48.5%)
7	179	137 (76.5%)
8	817	647 (79.2%)
9	193	133 (68.9%)
10	193	138 (71.5%)
11	485	295 (60.8%)
12	4571	3138 (68.7%)
13	305	207 (67.9%)
14	1024	249 (24.3%)
15	35	13 (37.1%)
16	822	675 (82.1%)
17	76	65 (85.5%)
18	536	340 (63.4%)
19	10	2 (20%)
20	3218	527 (16.4%)
21	386	291 (75.4%)
22	522	476 (91.2%)
23	363	91 (25.1%)
24	3155	2057 (65.2%)

A side-by-side comparison of the Relevance of the PQ applications and Phase 1 similar applications is shown below in **Figure 8**.

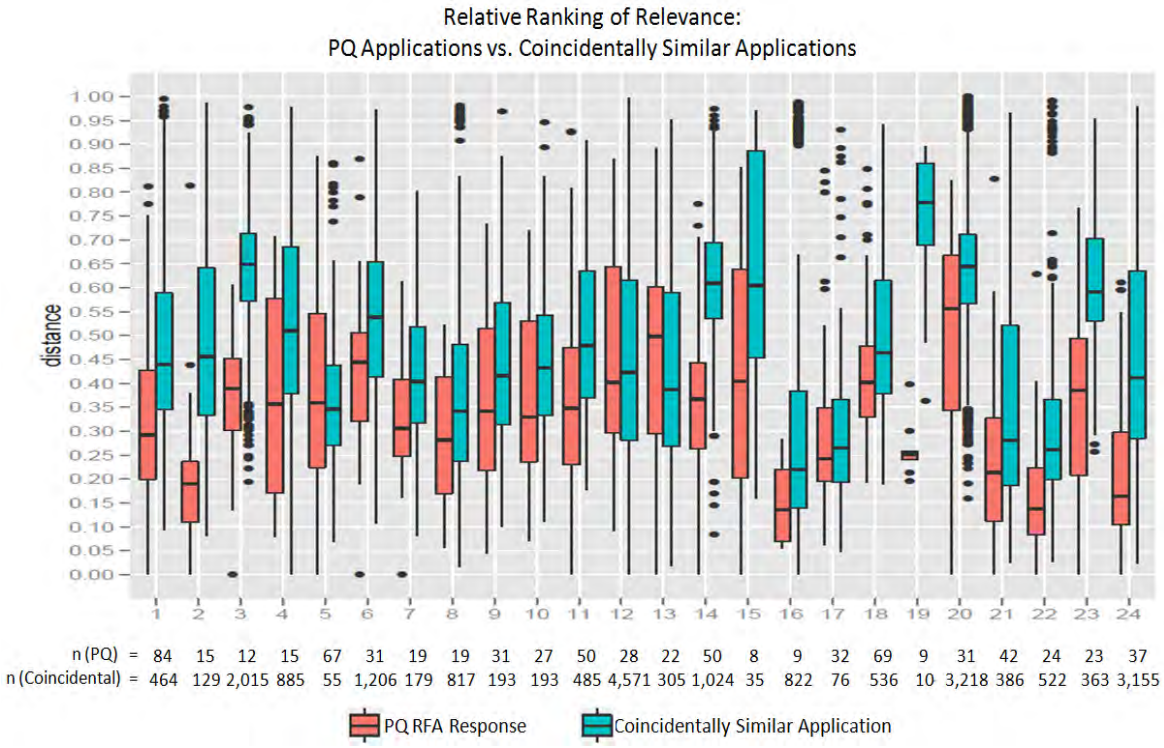

Figure 8. Relevance of RFA Responses and Similar Prior Grants to RFA Text.

Not surprisingly, the median "intentional" Relevance of the RFA responses is usually greater than that of the Phase 1 similar applications, but for PQs 5 and 13, the median Relevance of the Phase 1 similar applications noticeably exceeded that of the PQ applications.

3.2.3 Relevance Data Files

The details – comparison text, Relevance distance measurements, and a ranking of the merged set of both PQ and Phase 1 applications – are available in the **Relevance_Most_Responsive_Applications_with_Phase1_Merged.xslx** file. For each question, all PQ applications are listed in the file, along with their merged ranks. For each question, the application listing stops either after the last PQ application appears, or after the 100th application if all PQ applications are ranked higher than 100 on the merged list, leaving only Phase 1 similar applications not shown in the file[13]. Some of the outlier cases of non-responsive PQ applications have very high ranks on the merged list (especially in cases where there are many more Phase 1 applications than PQ applications for a given question). So the trailing set of least relevant PQ applications in those cases (e.g., PQ #1) will show large jumps in their ranks, since many Phase 1 applications are being left out. The **Relevance_Summary_by_Question.xslx** file shows the worst Relevance ranks for each question for both the PQ and Phase 1 similar application sets.

[13] All data is included in the side-by-side graph.

3.3 Combined Analysis

3.3.1 Relevance Definition and Relevance/ Focus Shift Quadrants

As a first step in analyzing the joint distributions of the Focus Shift and Relevance of the PQ applications, we now formally define ***Relevance*** as (1 – *Relevance distance)*, so that the value of 0 represents the least relevant and 1 the most relevant, and for both text metrics a numerical increase corresponds to a gain in a desirable application characteristic. Note that this defines the threshold for Relevance as 0.47 = 1-0.53). An application is relevant if Relevance ≥ 0.47. As before, an application is focus-shifted if Focus Shift ≥ 0.53

For the 702 PQ applications for which prior By-Self applications were found, we now have 2 sets of paired values (Focus Shift By-Self, Relevance), and (Focus Shift General, Relevance). For the remaining 52 applications we have only the (Focus Shift General, Relevance) pair. In this section, we examine graphic and statistical analysis of the distribution of these paired values, both overall and broken down by various factors, such as question number and PI career stage.

As shown below in **Figure 9**, in a scatterplot of the (Focus Shift, Relevance) values, the 2 thresholds define 4 quadrants in which a given application can be found:

- Neither Focus Shifted nor relevant, lower left quadrant, abbreviated as **
- Focus shifted but not relevant, lower right quadrant, abbreviated as Fs*
- Relevant but not Focus Shifted, upper left quadrant, abbreviated as *R
- Focus shifted and relevant, upper right quadrant, abbreviated as FsR

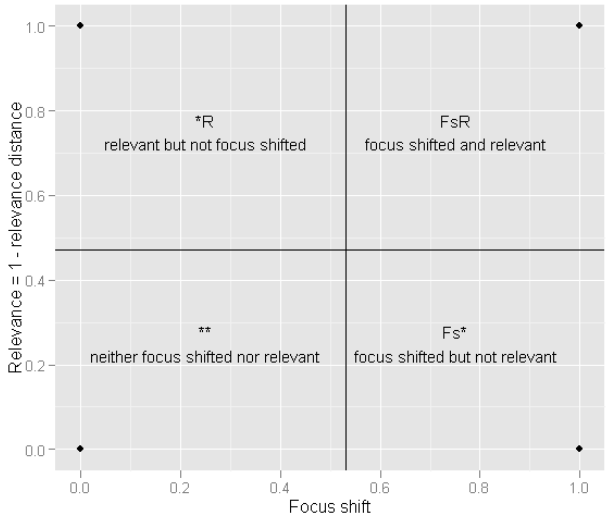

Figure 9. Focus Shift /Relevance Quadrants.

3.3.1 Overall and Per-Question Distributions of PQ Applications Over the FsR Quadrants

Table 9 shows the overall distribution of PQ applications across the Focus Shift/Relevance quadrants, using the Focus Shift by-self measurement. This table includes as 754 applications – those with no prior By-Self applications were classified into either the (**) or (*R) quadrants depending on whether they were relevant. **Table 10** shows the quadrant distribution, using the Focus Shift general measurement.

Table 11 (By-Self) and **Table 12** (General) show the distributions across the quadrants for the group of applications received for each Provocative Question.

Table 9. Quadrant Distribution, Using Focus Shift By-Self.

Focus shift by-self and Relevance classification	Description	PQ application count	Percentage of applications
FsR	Focus Shifted and Relevant	26	3.4%
*R	Relevant but not Focus Shifted	588	78.0%
Fs*	Focus Shifted but not Relevant	13	1.7%
**	neither Focus Shifted nor Relevant	127	16.8%

Table 10. Quadrant Distribution, Using Focus Shift General.

Focus shift general and Relevance classification	Description	PQ application count	Percentage of applications
FsR	Focus Shifted and Relevant	182	24.1%
*R	Relevant but not Focus Shifted	432	57.3%
Fs*	Focus Shifted but not Relevant	89	11.8%
**	neither Focus Shifted nor Relevant	51	6.8%

Table 11. Quadrant Distribution, By Question, Using Focus Shift By-Self.

Question #	Description	FsR	*R	Fs*	**
1	obesity & cancer	0.00	0.89	0.01	0.10
2	environmental risks	0.00	0.93	0.07	0.00
3	risk exposure	0.00	0.92	0.00	0.08
4	altering behaviors	0.00	0.67	0.00	0.33
5	off-label drugs	0.03	0.70	0.00	0.27
6	disease correlation	0.03	0.77	0.00	0.19
7	age dependence	0.00	0.89	0.00	0.11
8	tumor development	0.05	0.95	0.00	0.00
9	driver mutations	0.00	0.84	0.03	0.13
10	driver vs. passenger	0.00	0.74	0.04	0.22
11	alternative splicing	0.00	0.82	0.04	0.14
12	novel infectious agents	0.00	0.68	0.04	0.29
13	early detection	0.05	0.59	0.00	0.36
14	malignancy precursors	0.08	0.74	0.00	0.18
15	second primary cancers	0.13	0.50	0.13	0.25
16	metastases clinical significance	0.00	1.00	0.00	0.00

17	combination therapies	0.16	0.69	0.03	0.13
18	undruggable targets	0.04	0.78	0.01	0.16
19	chemo-only cures	0.00	1.00	0.00	0.00
20	immunotherapy biomarkers	0.00	0.39	0.00	0.61
21	resistance to radiotherapy	0.05	0.88	0.02	0.05
22	oncogene addiction	0.04	0.92	0.00	0.04
23	spontaneous regression	0.09	0.74	0.09	0.09
24	metastasis study techniques	0.08	0.81	0.00	0.11

Table 12. Quadrant Distribution, By Question, Using Focus Shift General.

Question #	Description	FsR	*R	Fs*	**
1	obesity & cancer	0.32	0.57	0.07	0.04
2	environmental risks	0.47	0.47	0.07	0.00
3	risk exposure	0.83	0.08	0.08	0.00
4	altering behaviors	0.40	0.27	0.27	0.07
5	off-label drugs	0.25	0.48	0.15	0.12
6	disease correlation	0.23	0.58	0.16	0.03
7	age dependence	0.26	0.63	0.11	0.00
8	tumor development	0.16	0.84	0.00	0.00
9	driver mutations	0.29	0.55	0.16	0.00
10	driver vs. passenger	0.15	0.59	0.22	0.04
11	alternative splicing	0.22	0.60	0.10	0.08
12	novel infectious agents	0.36	0.32	0.21	0.11
13	early detection	0.32	0.32	0.32	0.05
14	malignancy precursors	0.32	0.50	0.12	0.06
15	second primary cancers	0.25	0.38	0.13	0.25
16	metastases clinical significance	0.11	0.89	0.00	0.00
17	combination therapies	0.16	0.69	0.16	0.00
18	undruggable targets	0.23	0.59	0.13	0.04
19	chemo-only cures	0.22	0.78	0.00	0.00
20	immunotherapy biomarkers	0.10	0.29	0.16	0.45
21	resistance to radiotherapy	0.17	0.76	0.05	0.02
22	oncogene addiction	0.08	0.88	0.00	0.04
23	spontaneous regression	0.13	0.70	0.09	0.09
24	metastasis study techniques	0.05	0.84	0.03	0.08

3.3.1 Quadrant Scatterplot Analysis

Figure 10 shows the Relevance and Focus Shift plots for all PQ applications for which By-Self prior applications and peer review scoring data from the first round were available. Across all PQs, there were no clear "stand-out" applications that achieve the highest score for both Relevance and Focus Shift By-

Self; in many cases, applications scored high in one area, but in the lower to mid-range for the other. Interestingly, many of the applications with the "best" (score range 10) priority scores from peer review tended to have high Relevance to the RFA text, and were not Focus Shifted compared to prior applications. Exceptions to this trend were seen for PQs 1 and 5.

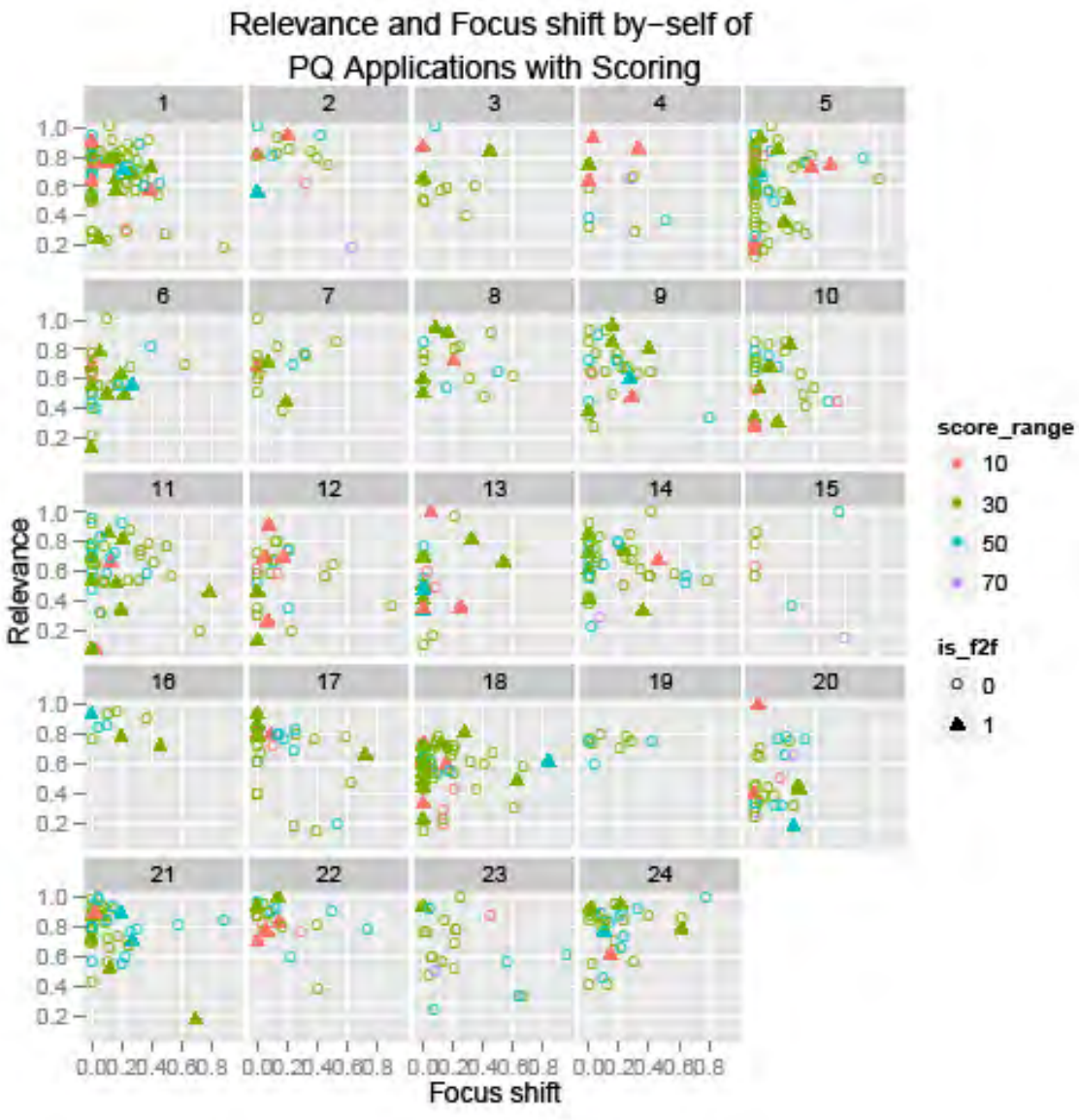

Figure 10. Relevance and Focus Shift By-Self of PQ Applications, By Question, With Scoring and F2F Status.

Figure 11 shows a similar analysis of the applications with scoring data from the first round, using the measurements of Focus Shift relative to the general subset. Here we see some definitive trends, such as the suggestion of a Relevance-Focus Shift tradeoff "boundary", either cutting down from left to right diagonally, particularly evident in PQs 1,6, 9, 11,17, and 18.

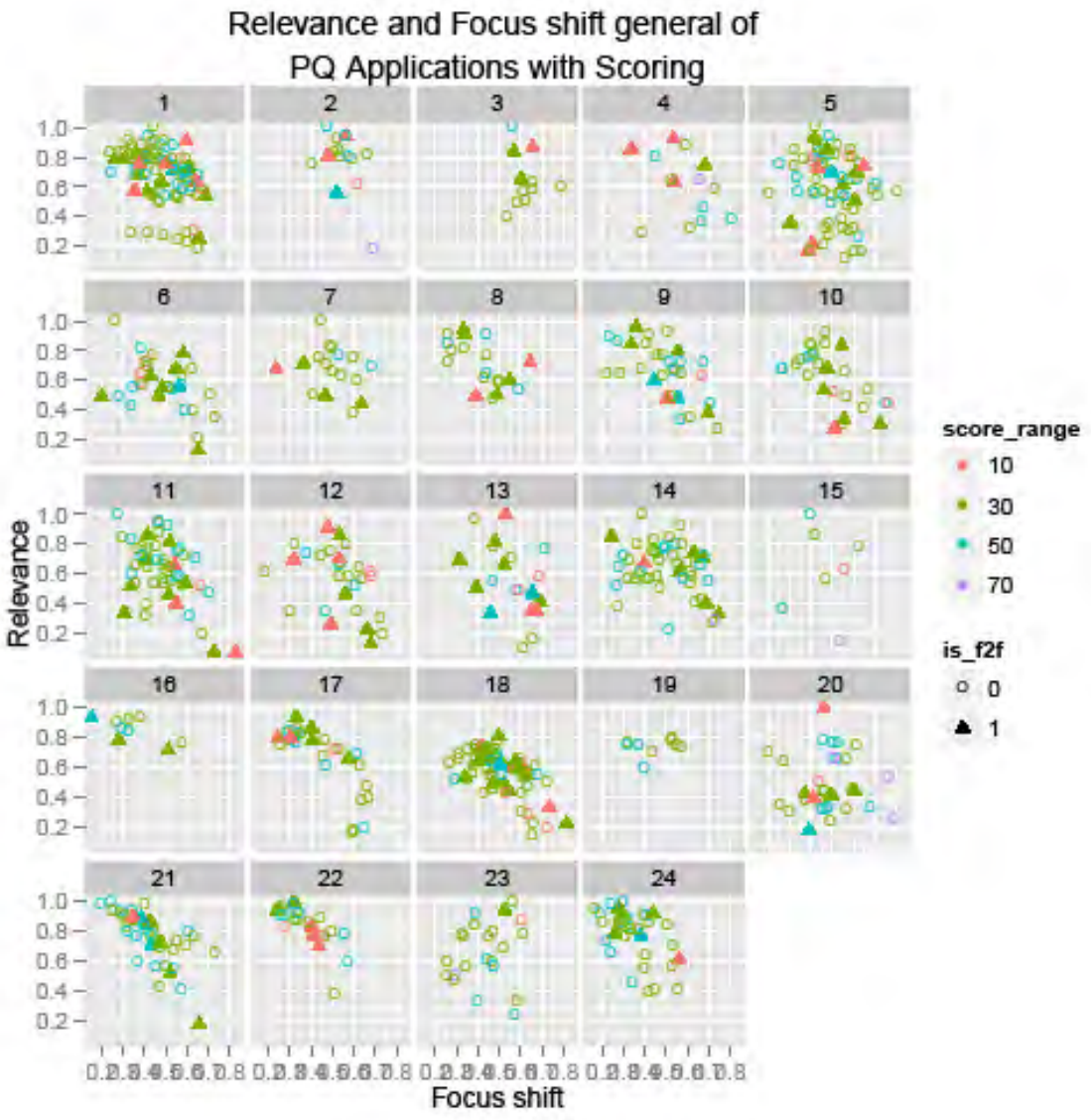

Figure 11. Relevance and Focus Shift General of PQ Applications, By Question, With Scoring and F2F Status.

As a supplemental graphics delivery, we also prepared a set of labeled, "zoomed-in" views for each of the panels above, one page per question, restricted to the 157 applications that advanced to the face-to-face review. Each data point is shown labeled with the application serial number, followed by its priority score in parenthesis. An example of one of these 24 pages, for PQ 8, is shown in **Figure 12**.

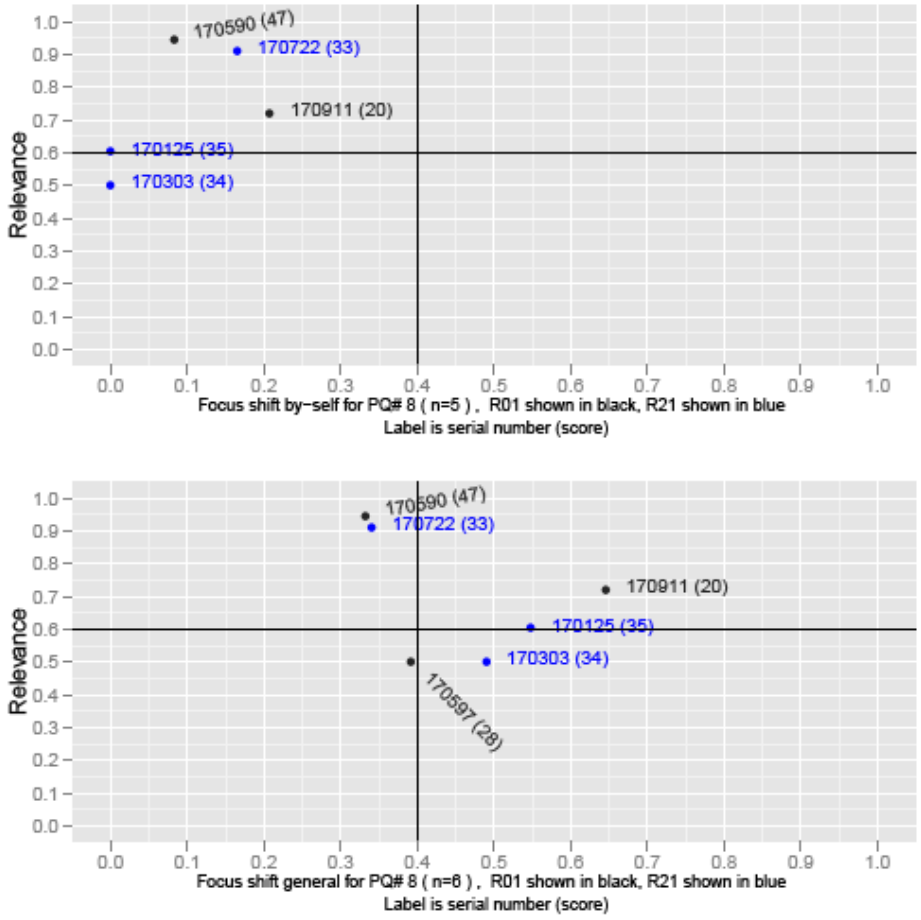

Figure 12. Relevance and Focus of Applications for PQ #8 Labeled by Serial Number and Priority Score.

3.3.1 Quadrant Analysis by Factors Characterizing the PQ Applicant Communities

We answered questions about how particular groups of investigators performed in their placement in four quadrants (with FsR being the best, ** the worst and Fs* / *R in the middle), with groups determined by question number, investigator degree category, investigator career stage, and the application mechanism (R01 or R21). We applied both graphic analysis, generating scatterplots color-coded by the factor levels, and statistical analysis, calculating unusually high or low representation in quadrants of interest (particularly FsR) based on a χ^2 test.

3.3.1.1 Quadrant Scatterplots

All graphs were generated at the question level, using both the By-Self and General Focus Shift measurements, for the 3 remaining factors (degree, career stage, and mechanism), resulting in 144 scatterplots (144 = 24 questions X 2 Focus Shift measurements X 3 factors). **Figures 13**, **14**, and **15,** below, show the pages generated for PQ 17.

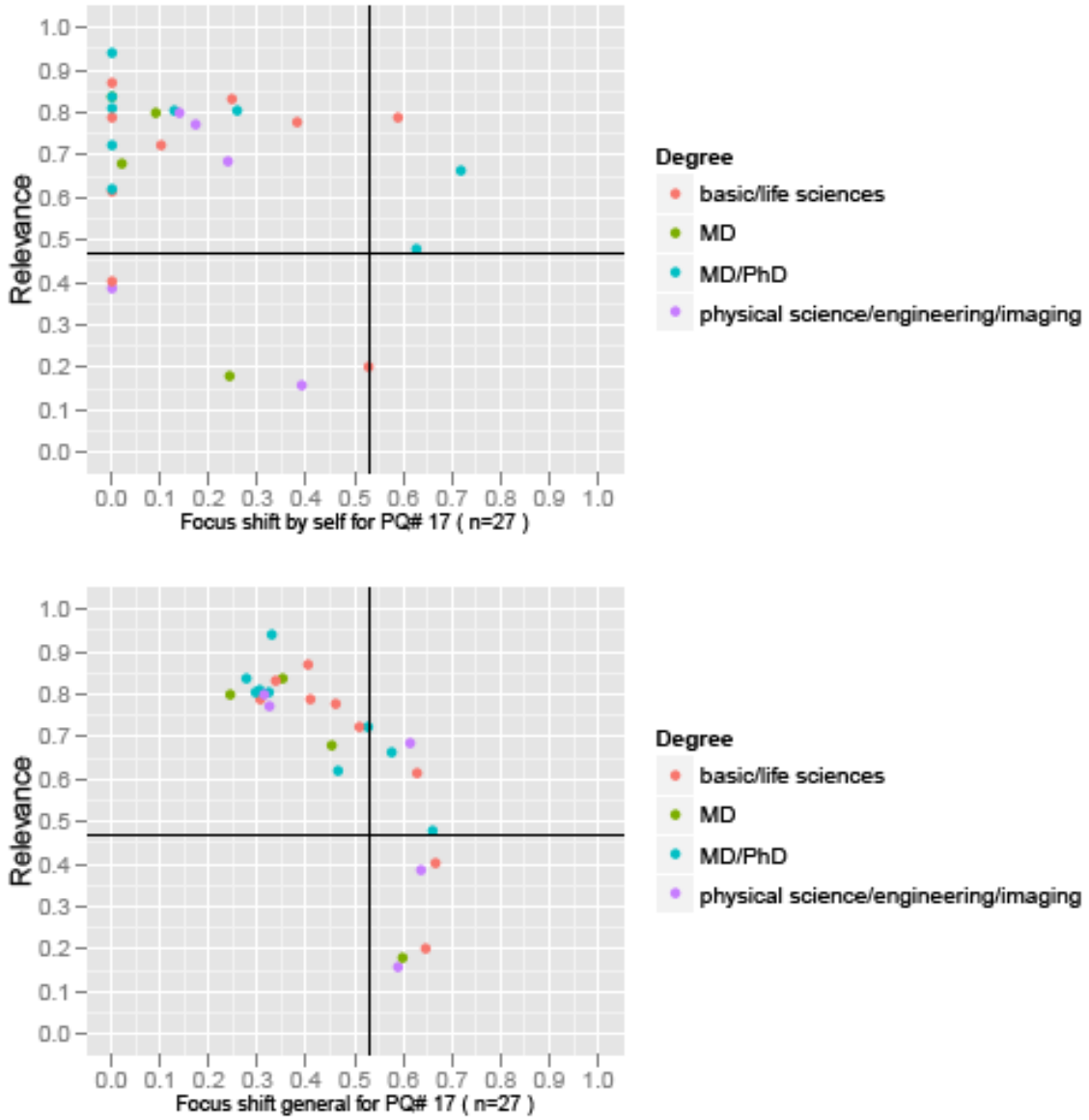

Figure 13. Relevance and Focus Shift of Applications for PQ #17 Labeled by Degree Category.

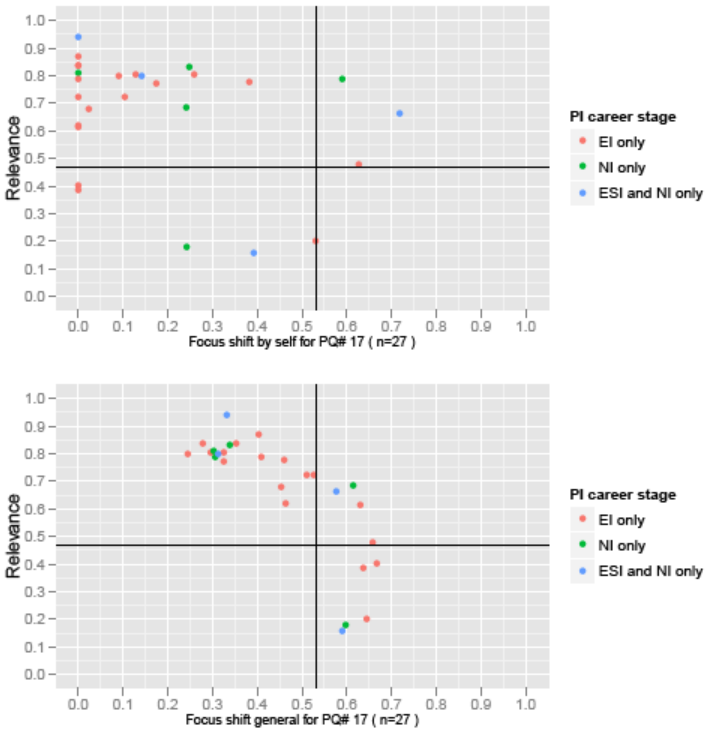

Figure 14. Relevance and Focus Shift of Applications for PQ #17 Labeled by PI Career Stage (EI = established investigator, NI = new investigator, ESI = early stage investigator).

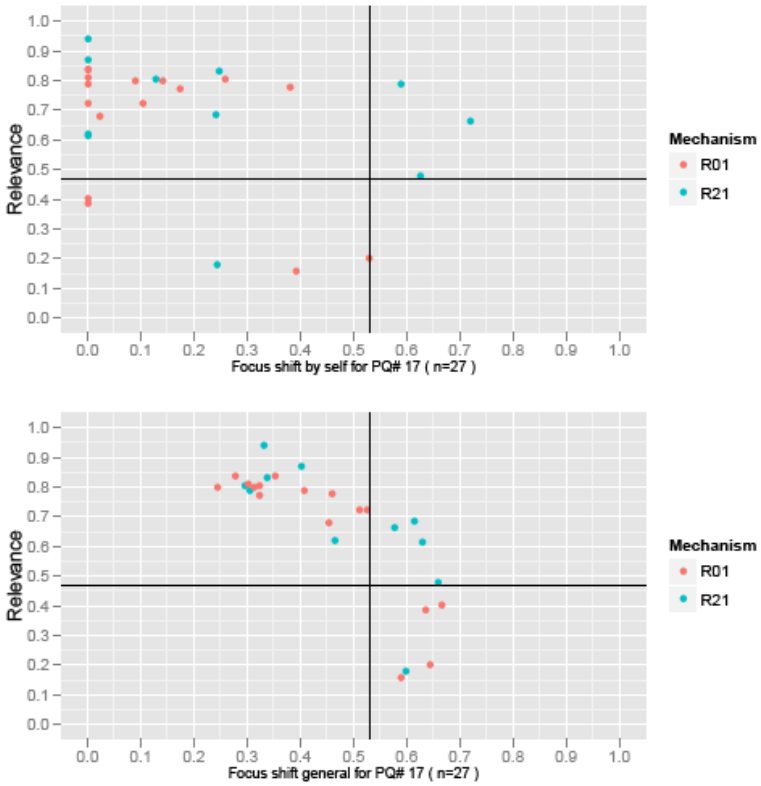

Figure 15. Relevance and Focus Shift of Applications for PQ #17 Labeled by Mechanism.

3.3.1.2 Quadrant χ^2 Analysis

To determine whether the distribution of applications across the Focus Shift/Relevance quadrants was correlated with degree, career stage, mechanism, or the PQ number, we performed 8 χ^2tests of independence based on contingency tables of sizes 4 x 4, 3 x 4, 2 x 4, and 24 x 4, respectively. Since some expected cell counts were less than 5, we used Monte Carlo simulated p values. We used standardized residuals to identify cells (defined by a particular quadrant and a particular value of one of the factors) for which the number of applications was either significantly higher or lower than would be expected if the factor and the quadrant distribution were independent. A standardized residual greater than 2 in absolute value was used as a threshold for a significant overrepresentation or underrepresentation in a given cell.[14]

Tables 13, **14**, and **15,** below, show the 3 tests for which significant ($p < 0.001$) results were obtained.

In **Table 13**, where the Focus Shift measurement is relative to the By-Self set of prior applications, we see that established investigators had significantly fewer focus-shifted and relevant applications than expected (given the overall numbers of EI-only cases and FsR applications), and that applications with new and early-stage investigators had significantly more applications in this "best case" quadrant. However, new investigators, by themselves, had more trouble achieving Relevance.

Table 13. Focus shift by-self: standardized residuals for the Career stage X Quadrantχ^2 test. Values > 2 or < -2 indicate significant over or under representation in a given Focus Shift/Relevance quadrant for either EI = established investigators, ESI and NI = a combination of new and early-stage investigators, or NI = new investigators.

Career stage	FsR	*R	Fs*	**
EI only	-4.5	2.3	-2.0	0.3
ESI and NI only	2.8	0.8	-0.8	-2.0
NI only	2.8	-3.4	3.0	1.4

Focus Shifting **Table 14**, we compare focus-shift By-Self with the RFA question number, and we see several PQ RFAs with significantly higher or lower representation in various Focus Shift/Relevance quadrants. Most notably, PQs 15, 17, and 23 did particularly well in achieving Focus Shift, although for PQ 15, this may have been at the cost of higher Relevance. PQs 5, 13, and 20 had a higher than expected number of applications that were measured as neither Focus Shifted nor relevant.

Table 15 compares focus-shift General with the RFA question number, and we see quite a different set of results, with PQs 2,3, and 13 now showing higher representation in the focus shift General quadrants, although PQ 20 still stands out as overrepresented in the (**) quadrant.

[14] cf. Agresti, A. (2002). *Categorical Data Analysis*. Wiley-Interscience.

Table 14. Focus Shift By-Self: Standardized Residuals for the PQ # Stage X Quadrant χ^2 Test.

PQ	FsR	*R	Fs*	**
1	-1.8	2.7	-0.4	-1.9
2	-0.7	1.4	1.5	-1.8
3	-0.7	1.2	-0.5	-0.8
4	-0.7	-1.1	-0.5	1.7
5	-0.2	-1.6	-1.1	2.3
6	-0.1	-0.1	-0.8	0.4
7	-0.8	1.2	-0.6	-0.7
8	0.4	1.8	-0.6	-2.0
9	-1.1	0.8	0.7	-0.6
10	-1.0	-0.5	0.8	0.8
11	-1.4	0.7	1.3	-0.6
12	-1.0	-1.3	0.8	1.7
13	0.3	-2.2	-0.6	2.5
14	1.8	-0.7	-1.0	0.2
15	1.4	-1.9	2.4	0.6
16	-0.6	1.6	-0.4	-1.4
17	3.9	-1.3	0.6	-0.7
18	0.4	0.1	-0.2	-0.2
19	-0.6	1.6	-0.4	-1.4
20	-1.1	-5.4	-0.8	6.8
21	0.5	1.6	0.3	-2.2
22	0.2	1.6	-0.7	-1.7
23	1.4	-0.5	2.6	-1.1
24	1.6	0.5	-0.8	-1.0

Table 15. Focus General: Standardized Residuals for the PQ # Stage X Quadrant χ^2 Test.

PQ	FsR	*R	Fs*	**
1	1.8	0.0	-1.4	-1.2
2	2.1	-0.8	-0.6	-1.1
3	4.8	-3.5	-0.4	-0.9
4	1.5	-2.4	1.8	0.0
5	0.2	-1.7	0.8	1.8
6	-0.2	0.1	0.8	-0.8
7	0.2	0.5	-0.2	-1.2
8	-0.9	2.4	-1.6	-1.2
9	0.7	-0.3	0.8	-1.5
10	-1.2	0.2	1.7	-0.6
11	-0.4	0.4	-0.4	0.4
12	1.5	-2.7	1.6	0.8
13	0.9	-2.5	3.0	-0.4
14	1.3	-1.1	0.0	-0.2
15	0.1	-1.1	0.1	2.1
16	-0.9	1.9	-1.1	-0.8
17	-1.2	1.3	0.7	-1.6
18	-0.2	0.4	0.3	-0.8
19	-0.1	1.2	-1.1	-0.8
20	-1.9	-3.2	0.8	8.7
21	-1.2	2.5	-1.5	-1.2
22	-1.8	3.0	-1.8	-0.5
23	-1.3	1.2	-0.5	0.4
24	-2.7	3.3	-1.8	0.3

4.0 Other Relationships Between Focus Shift, Relevance and Application/ Applicant Characteristics

In this section, we examine how the Focus Shift and Relevance measurements are correlated with the results of the review panel's manual evaluation of an application, whether particular questions attracted new and early-stage investigators, and whether particular questions attracted a more diverse group of investigators. More generally, this section describes applicant and application characteristics associated with higher funding probability and higher scores for Focus Shift or Relevance.

A unifying hypothesis in this analysis was that the Focus Shift and Relevance measurements represent new data that is not already incorporated into the evaluation process by some other metric. If the hypothesis is confirmed, this would suggest these measurements (or a refined version of them) may be useful to inform the grant evaluation or program evaluation process.

4.1 Focus Shift / Relevance and F2F or Award Status: the "Naïve Predictor"

Although many factors enter into the evaluation process, we studied the question (with a deliberate degree of naiveté) of whether the Focus Shift and Relevance measurements, on their own, could be used as predictors of whether applications would advance to the face-to-face evaluation phase, or would ultimately be awarded. We constructed a "naïve" predictor by computing a weighted pseudo-L_2 norm called an "application radius" on the 3D vector formed by the Focus Shift By-Self, Focus Shift General, and Relevance measurements. We predicted which applications would have F2F status, or be awarded based on whether the application radius exceeded a pre-selected threshold.

A preliminary version of the naïve predictor, with award status as the predicted outcome, and using arbitrary fixed weights of 0.3, 0.2, and 0.5 for Focus Shift By-Self, Focus Shift General, and Relevance, had 17% sensitivity (recall) and 87% specificity. This placement is only slightly above the worst-predictor line on a ROC curve.

To see if improvement was possible, we switched to predicting F2F status using a set of 18 different weight triplets. We used ROC curves to plot the predictive success for the full range of thresholds for each of the 18 weight sets applied to 3 different data subsets: the full set of applications, the R01 applications, and the R21 applications. None of the predictors were found to be of high quality.

- The best predictor for all applications used the weights (0.1, 0.7, 0.2) and gave AUC = 0.53, specificity = 64% sensitivity = 44%
- The best predictor for R01's used the Mahalanobis weights (29, 59, 25) and gave AUC = 0.56, specificity = 52% sensitivity = 64%
- The best predictor for R21's used the weights (0.8, 0.1, 0.1) and gave AUC = 0.59, specificity = 32% sensitivity = 86%

The ROC curve for the best R01 predictor is shown below in **Figure 16**.

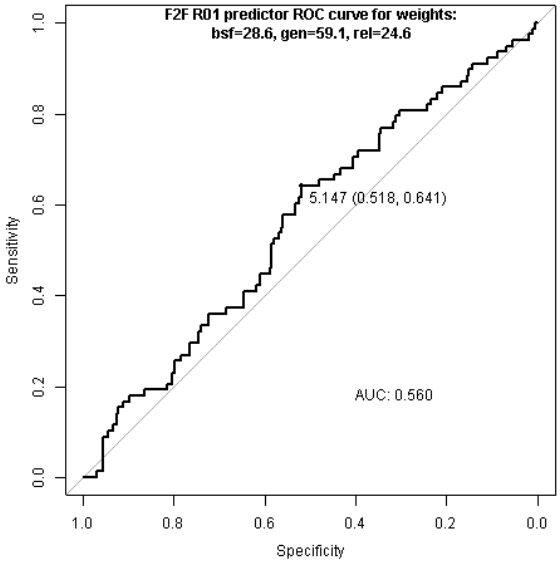

Figure 16. ROC Curve for Binary Classifier Used to Predict F2F Status, Using Weights of 28.6, 59.1, and 24.6.

The conclusion is that the text measurements of Focus Shift and Relevance by themselves, using a norm/threshold model, are not good predictors of funding or F2F status; this may be viewed as evidence supporting the hypothesis that these are new measurements which are currently not directly incorporated into the existing evaluation process.

4.2 RFA Questions and Applicant Characteristics

We examined two questions regarding the investigator populations associated with each PQ RFA: (1) Do particular questions attracted a higher proportion of new and early stage investigators? (2) Do particular questions attracted a more diverse group of investigators?

In answer to question (1), **Figure 17** shows the proportions of the 3 distinct career stage combinations observed in the 754 PQ applications (only early stage and new investigators, only new investigators, and only established investigators), computed as a proportion or probability at the question level. **Table 16** summarizes the questions for which the proportion of ESI and/or NI investigators was noticeably higher or lower than average.

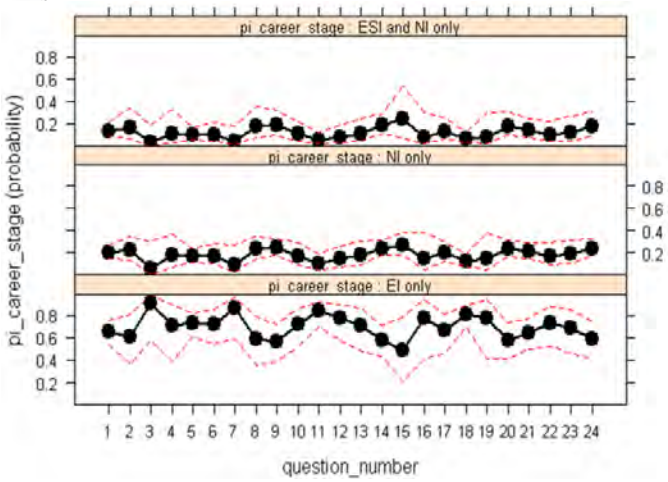

Figure 17. Proportions of the 3 observed career stage patterns in the 754 PQ applications aggregated by question number. The dashed lines represent 95% confidence intervals, assuming some hypothetical variation in the EI, NI, and ESI from the observed values.

Table 16. Questions with Higher or Lower Proportions of New or Early Stage Investigators.

Provocative question number	Question topic	Association with higher/lower probability of early-stage or new investigators
2	environmental risks	higher probability of ESI/NI
8	tumor development	higher probability of ESI/NI
9	driver mutations	higher probability of ESI/NI
14	malignancy precursors	higher probability of ESI/NI
15	second primary cancers	higher probability of ESI/NI
20	immunotherapy biomarkers	higher probability of ESI/NI
24	metastasis study techniques	higher probability of ESI/NI
3	risk exposure	lower probability of ESI/NI
7	age dependence	lower probability of ESI/NI
11	alternative splicing	lower probability of ESI/NI

To answer question (2), we defined a rough measure of investigator diversity using the Gini index of the distribution of application counts over the investigator's degree categories. The Gini index measures the extent to which applications are concentrated in a few (or one) degree category, rather than being distributed over a larger set of categories. The four degree categories were: (a) basic/life sciences, (b) MD, (c) MD/Ph.D., and (d) physical science/engineering/imaging. If all applications for a given question had PIs with a single degree category (the least diversity), the Gini index would be 0.75[15]. Conversely, if the applications for a given question were equally divided among the four degree categories (the most diversity) the Gini index would be 0. Therefore, lower values on the Gini index indicate questions with a more diverse pool of applicants.

[15] For just two categories, the maximum Gini index is 0.5, the maximum (least diverse) Gini index approaches 1 from below as the number of categories increases to infinity.

Figure18, below, shows the computed Gini index for each question. Question 13 is placed on the left, since in all regression models (discussed in sections 4.3 and 4.4), PQ13 was selected as the reference level against which other questions were compared for their effect on outcome variables. Question 13 was selected as the reference level because it had an overrepresentation of applications in the ** quadrant (neither Focus Shifted nor relevant), but not the extreme overrepresentation in that quadrant exhibited by PQ #20.

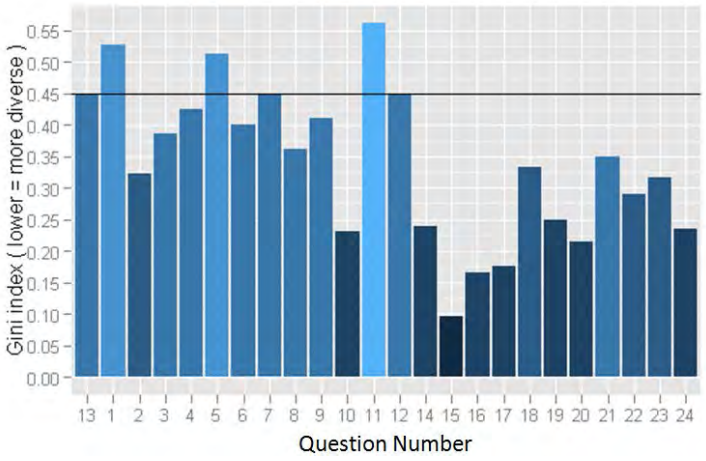

Figure 18. Gini index for each question. Lower values indicate higher diversity in PI degree categories.

Table 17 summarizes the noticeable outlier cases from **Figure 20**. There was also some weak evidence that an increase in Focus Shift General was associated with lower PI degree diversity (p = 0.068).

Table 17. Questions with Particularly High or Low Diversity in PI Degree Category.

Provocative question number	Question topic	Diversity of investigators as measured by the Gini index of the counts in each degree category
15	second primary cancers	high
16	metastases clinical significance	high
17	new drug testing	high
1	obesity & cancer	low
5	off-label drugs	low
11	alternative splicing	low

4.3 Scoring and Funding Correlations with Text Measurements and Applicant/Application Characteristics

We ran a series of linear and logistic models that provided the results presented in this section and Section 4.4. These models were based on the list of variables shown below in **Table 18**. In this section, we focus on the relationships between the scoring variables and the text measurement, and the relationship between the binary outcome of funding (Yes/No) and all other variables as inputs.

Table 18. Modeling Variables with Typical Values and Reference Levels.

Variable	Typical values	Reference level
question_number	1,2,3,…24	13
Is_F2F	1(Yes), 0 (No)	0
funded	1,0	0
relevance	decimal in range 0 to 1	
focus_shift_by_self	decimal in range 0 to 1	
focus_shift_general	decimal in range 0 to 1	
pi_career_stage	EI only, NI only, ESI and NI only	EI only
deg_cat	basic/life sciences, behavioral, epidemiology, MD, MD/PhD, physical science/engineering/imaging	basic/life sciences
ac	R01, R21	R01
deg_fos_prefix	61 distinct values including: medic unknown immun chemi pysch carci bioin bioch molec patho micro healt physi pharm other anima	unknown
org_state	46 distinct values including: TX CA PA NY SC NC OH	CA
primary_referral	CRCHD CSSI DCB DCCPS DCP DCTD unknown	unknown
is_multi_pi	1,0	0
approach	number in range 1.5 to 8.3, mean 4.6	
environment	number in range 1.0 to 6.7, mean 2.1	
innovation	number in range 1.0 to 8.0, mean 3.4	
investigators	number in range 1.0 to 7.3, mean 2.4	
priority_score	number in range 11 to 80, mean 41.5	

Before running the models, we examined the correlations between the criteria scores (innovation, environment, approach, investigators, and significance), the final priority score, and the three text measurements (Focus Shift By Self, Focus Shift General, and Relevance). A full tableau of all possible pairwise scatter plots is shown below in **Figure 19**. Visually, many of the criteria scores appear to be somewhat correlated with each other and with the final priority score, but there appears to be little correlation among any of these scores and the Focus Shift or Relevance measurements. There is also a suggestion of an inverse correlation: increasing Focus Shift General appears to be correlated with decreasing Relevance. **Table 19** confirms these visual observations, using the computed Kendall correlations between the variables.

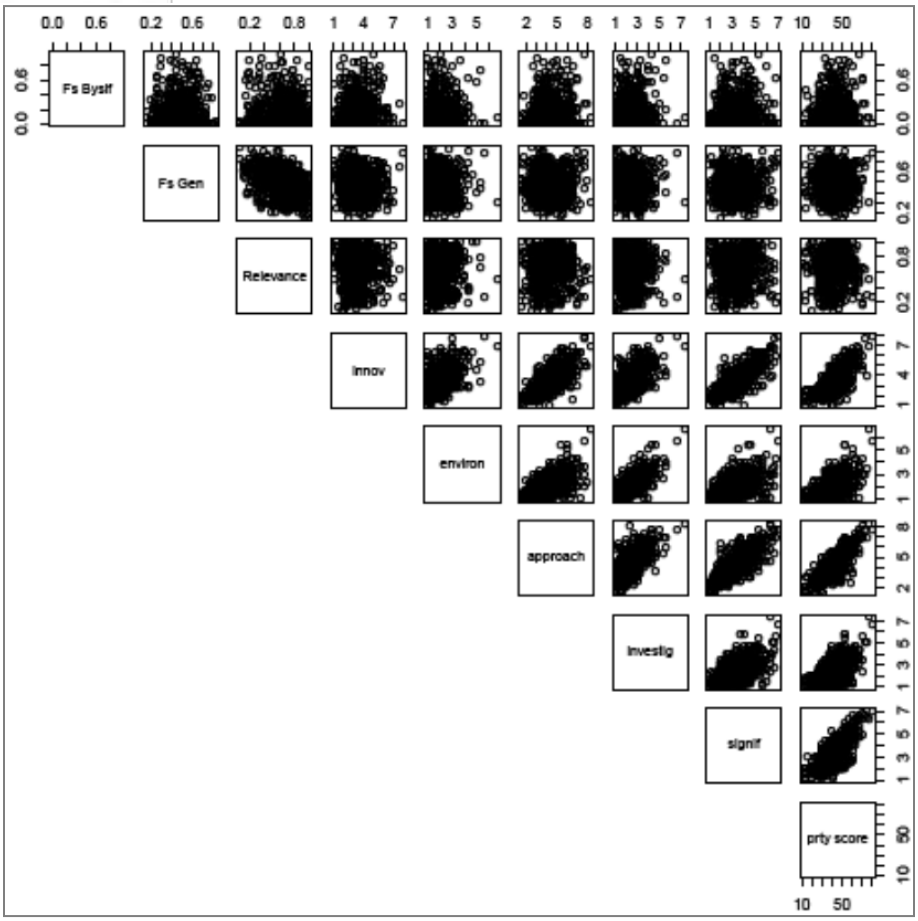

Figure 19. Scatterplots of Scoring Variables and Text Metrics.

Table 19. Kendall Correlation Coefficients of Scoring Variables and Text Metrics.

	focus shift general	relevance	innovation	environment	approach	investigators	significance	priority score
focus shift by self	-0.03	0.03	0.07	0.05	0.10	0.08	0.06	0.07
	focus shift general	-0.30	-0.04	-0.03	-0.02	-0.03	-0.01	-0.03
		relevance	0.02	0.02	0.03	0.01	0.01	0.03
			innovation	0.36	0.54	0.41	0.57	0.53
				environment	0.41	0.57	0.35	0.32
					approach	0.48	0.59	0.62
		large correlation				investigators	0.41	0.39
		medium correlation					significance	0.57

We then constructed a single-stage logistic regression model with funding status (Yes/No) as the output variable and all other variables listed in **Table 18** as inputs. This model found the significant variables

affecting the predicted probability of funding were DCCPS referral, having the "ESI and NI only" PI career stage classification, and, as the most significant variable, priority score.

We started to explore construction of a multi-stage or hierarchical model, but did not complete this analysis, although there was some evidence that the criteria scores for approach and innovation were significant factors (the most significant factor contributing to the priority score was F2F status and we did not construct a model of the F2F status outcome).

In all of these models, neither Focus Shift nor Relevance appeared as significant variables related to the funding decision. At a very relaxed level of confidence ($\alpha = 0.15$), we would report lower odds of funding for applications with a higher value of Focus Shift By Self.

Generally, these results support the hypothesis that Focus Shift and Relevance are independent measurements that could be considered for separate consideration in the review process.

4.4 Text Measurements Correlations with and Applicant/Application Characteristics

Finally, we constructed linear regression models to determine the significant associations between the application and applicant characteristics listed in **Table 18** and the measurements of Focus Shift and Relevance. The scoring variables were included, but as suggested by the correlation analysis, there was no significant association found between scores and text measurements.

The significant associations that were found are outlined below:

Significant association with ***higher Focus Shift By Self*** scores for

- NI only and ESI/NI only
- PQ #23 (spontaneous regression)
- Higher approach scores

Significant association with ***higher Focus Shift General*** scores for

- PQ #3 (risk exposure)
- PQ #13 (early detection)

Significant association with ***lower Focus Shift General*** scores for

- PQ #8 (tumor development)
- PQ #16 (metastases clinical significance)
- PQ #21 (resistance to radiotherapy)
- PQ #22 (oncogene addiction)
- PQ #24 (metastasis study techniques)

Significant association with higher ***Relevance*** scores for

- PQ # 16, 21, 22, 24

Significant association with ***lower Relevance*** scores for

- PQ #5 (off-label drugs)

PQ #6 (disease correlation)
PQ #12 (novel infectious agents)
PQ #18 (undruggable targets)
PQ #20 (immunotherapy biomarkers)
MD degrees
CRCHD, DCB, and DCCPS referrals

5.0 Conclusions

This work represents a first step toward the use of automated text mining algorithms to inform the grant evaluation process. Our results indicate that Focus Shift and Relevance values are attributes of the grant application that are currently not directly incorporated into the existing evaluation process. The primary limitation of our current approach to calculate these two quantities is that when two bodies of text are found to be similar, it may represent a similarity of background and stage setting rather than a similarity of the experimental approach. Generally, Focus Shift calculations were found to be more accurate than Relevance in terms of their agreement with with manual assessment of the scientific similarity between documents. Re-examining the choice of text used for analysis is likely to show promise in improving the confidence in the meaning of both Focus Shift and Relevance scores; both measurements may be improved by the inclusion of the specific aims section of grant applications. Another important next step is to use a more sophisticated text mining approach that accounts for symtantic relationships within the documents. Finally, two critical next steps are establishing a comparison group for several of these analyses and carrying out additional manual review by subject matter experts.

Appendix 1. Subdivision of PQ Applications by Extent of Focus Shift

Here we present supporting data for the discussion in section 3.1.2.

Table A1. Total Potentially "Repurposed" PQ Grant Applications

Question	Total Applications	Number Repurposed	Percent Repurposed
1. obesity & cancer	84	33	39%
2. environmental risks	15	4	27%
3. risk exposure	12	6	50%
4. altering behaviors	15	6	40%
5. off-label drugs	67	37	55%
6. disease correlation	31	15	48%
7. age dependence	19	8	42%
8. tumor development	19	6	32%
9. driver mutations	31	10	32%
10. driver vs. passenger	27	12	44%
11. alternative splicing	50	17	34%
12. novel infectious agents	28	8	29%
13. early detection	22	13	59%
14. malignancy precursors	50	25	50%
15. second primary cancers	8	4	50%
16. metastases clinical significance	9	3	33%
17. combination therapies	32	12	38%
18. undruggable targets	69	34	49%
19. chemo-only cures	9	4	44%
20. immunotherapy biomarkers	31	13	42%
21. resistance to radiotherapy	42	18	43%
22. oncogene addiction	24	9	38%
23. spontaneous regression	23	6	26%
24. metastasis study techniques	37	8	22%

Table A2. PQ Grant Applications Potentially "Repurposed" from Prior Unfunded Grant Applications

Question	Total Applications	Number Repurposed	Percent Repurposed
1. obesity & cancer	84	20	24%
2. environmental risks	15	3	20%
3. risk exposure	12	4	33%
4. altering behaviors	15	5	33%
5. off-label drugs	67	21	31%
6. disease correlation	31	11	35%
7. age dependence	19	4	21%
8. tumor development	19	2	11%
9. driver mutations	31	5	16%
10. driver vs. passenger	27	11	41%
11. alternative splicing	50	8	16%
12. novel infectious agents	28	6	21%
13. early detection	22	9	41%
14. malignancy precursors	50	19	38%
15. second primary cancers	8	3	38%
16. metastases clinical significance	9	2	22%
17. combination therapies	32	6	19%
18. undruggable targets	69	17	25%
19. chemo-only cures	9	2	22%
20. immunotherapy biomarkers	31	7	23%
21. resistance to radiotherapy	42	9	21%
22. oncogene addiction	24	3	13%
23. spontaneous regression	23	3	13%
24. metastasis study techniques	37	6	16%

Table A3. PQ Grant Applications Potentially "Repurposed" from Prior Funded Grants with Publications

Question	Total Applications	Number Repurposed	Percent Repurposed
1. obesity & cancer	84	9	11%
2. environmental risks	15	0	0%
3. risk exposure	12	2	17%
4. altering behaviors	15	0	0%
5. off-label drugs	67	13	19%
6. disease correlation	31	2	6%
7. age dependence	19	4	21%
8. tumor development	19	2	11%
9. driver mutations	31	4	13%
10. driver vs. passenger	27	1	4%
11. alternative splicing	50	9	18%
12. novel infectious agents	28	0	0%
13. early detection	22	2	9%
14. malignancy precursors	50	4	8%
15. second primary cancers	8	1	13%
16. metastases clinical significance	9	1	11%
17. combination therapies	32	4	13%
18. undruggable targets	69	13	19%
19. chemo-only cures	9	2	22%
20. immunotherapy biomarkers	31	4	13%
21. resistance to radiotherapy	42	4	10%
22. oncogene addiction	24	3	13%
23. spontaneous regression	23	3	13%
24. metastasis study techniques	37	1	3%