Methods for Identifying Misvalued Work Relative Value Units (RVUs)

One of the two key components of this project is the development of methods by which HCFA could identify CPT codes in the Medicare fee schedule (MFS) that may have been assigned inappropriate work RVUs, e.g., overvalued CPT codes, undervalued CPT codes, and work RVU anomalies. Additionally, we are interested in developing methods that could be used to identify services whose work RVUs may have changed since they were originally developed or last revised, or to provide quantitative data to clinicians in the development of work RVUs.

As a first step, we evaluated methods used by HCFA and the RUC during the initial five year review. We also conducted a review of the published literature looking for methods that had been used by researchers to describe the work of physicians in providing medical care or to measure elements of work, e.g., intra-service time. Methods were explored and dismissed, if we were unable to locate representative databases or to develop reasonable primary data collection strategies to support the method. For example, we explored the use of physician appointment scheduling software to identify services whose estimated service times underlying current work RVUs might different significantly from average appointment times built into or abstracted from the scheduling software. During our evaluation of this method, we learned that most scheduling software is customized to the physician requiring each practice to enter in their “normal” estimates of minutes required for each service physicians within the practice provide. We were unable to locate scheduling software that provided physicians with “normative” time estimates.

The result of our efforts is a set of seven methods that could potentially be used by HCFA to identify misvalued services, and that focuses upon different components of the work RVU: total work; time components of total work, i.e., total service time, pre- and intra-service time, and post-operative visits; and work per unit time. Several methods allow for the evaluation of multiple components simultaneously.
Methods for Identifying Misvalued Work Relative Value Units (RVUs)

For each of the proposed methods, we provide the following information:

- a brief overview of the method;
- a background discussion highlighting the theoretical or conceptual underpinnings of the proposed method;
- a detailed description of necessary data;
- a detailed description of the method;
- an illustrative example of the method;
- the projected time frame to conduct the proposed method;
- strengths and weaknesses of the proposed method; and
- the likely response from key stakeholders.

Prior to providing this information, we briefly summarize each method.

To identify services whose total work (as well as pre-service work) may be inappropriately valued, we present two methods: clinical panels and Rasch paired comparisons. Section 2.1 provides a detailed discussion of the role of clinical panels in evaluating total physician work and concentrates on issues related to panel composition and mode of deliberation using the technique of magnitude estimation. A review of HCFA’s previous use of panels to establish and/or evaluate work RVUs is provided and recommended strategies for conducting future panels is provided. Section 2.2 introduces an alternative, simplified method for identifying misvalued procedures. The Rasch paired comparison method identifies misvalued CPT codes in terms of either total or intra-service work based on a small-group panel comparison of codes within clinical families. The objective is to identify statistical outliers that appear to be either misaligned or compressed in terms of overall physician work effort employing a simpler approach than magnitude estimation.

Time is a significant component in determining physician work. The Harvard RBRVS study used time as a fundamental component in determining the relative differences in work and the RUC has incorporated time into its methodology for determining new or revised work estimates. In theory, if the underlying time estimates are appropriate for all services, one should be able to multiply the time estimates associated with current work RVUs by the volume of services a
physician provides within a specified time period to generate a total service time period that is reflective of the actual number of clinical hours the physician spent providing the services. To assess the appropriateness of total service time underlying current work estimates, in Section 2.3 we propose the use of a physician-level clinical profile database that contains estimates of total available clinical time as well as estimates of total volume of services provided during that time period at the CPT code level. Thus, this method uses objectively collected volume of service estimates and current work RVU time estimates to evaluate the reasonableness of the total service time estimates relative to estimates of available clinical time.

It is generally acknowledged that physician work has three main components — pre-service, intra-service, and post-service work. Time serves as an important dimension for each of these components with intra-service time being a major component of all services. Intra-service time is defined as the face-to-face time for nonsurgical procedures or the skin-to-skin time for surgical procedures. The Harvard study demonstrated a strong positive correlation between intra-service time and work, thus a significant change in time is likely to lead to a clinically meaningful change in work. Historically, time has been captured through physician surveys. However, current intra-service time estimates provided by the RUC appear to be systematically different, and typically higher, than the times measured by the original Harvard study. On average, the RUC intra-service time estimates for the same services are approximately 14 percent higher than those measured by Harvard, while total service times are 20 percent higher. This finding suggests that a more reliable database is needed to provide actual times for the intra-service component of care. In Section 2.4, four alternative objective data sources are discussed for the identification of services whose intra-service times may be misvalued: anesthesia times estimated from Medicare claims data; operative times obtained from a data vendor; operative times collected from a panel of hospitals; and intra-service times collected through direct observation.

Under Medicare’s RBRVS fee schedule, a single fee is paid for all necessary services provided by the surgeon before, during, and after the surgery¹. The pre-operative period for major

¹ Separate payment windows exist for major surgery, minor surgery, and nonincisional procedures.
Methods for Identifying Misvalued Work Relative Value Units (RVUs)

surgery starts on the day before the surgery and includes all visits, in or out of the hospital. The intra-operative period comprises the normal and necessary services that are required to perform the surgery. The post-operative period is 90 days for major surgeries and all related evaluation and management services are considered part of the post-operative period. Researchers at Harvard asked physicians to provide judgments about the typical number and duration of pre- and post-operative visits that occur during the global surgical periods. In developing work RVUs for new or revised CPT codes, the RUC also surveys physicians to obtain estimates of the number and total time of visits that occur during the global surgical period. The method described in Section 2.5 uses Medicare claims data to identify services whose number of pre- and post-operative hospital visits provided during the global surgical period may be misvalued given current lengths of stay and proportion of same day surgery cases.

The method described in Section 2.6 uses Medicare claims data to identify services with potentially misvalued work RVUs by analyzing changes in site-of-service, total frequency, and specialty mix of services over time. As such, it builds upon analyses conducted by the RUC during the initial five year review; the RUC analyzed trends in the frequency and site-of-service for services provided between 1992 and 1994 with the intent of identifying potentially overvalued services. It believed that a high rate of increase in frequency and shift in site-of-service could indicate that the service was becoming more commonly furnished and that the work involved may have decreased.

A primary objective of this method is to identify services whose locus has shifted from one setting to another, e.g., from the inpatient to outpatient setting, since the development of their work RVUs. The reason being is that the nature and scope of pre- and post-service work is likely to change as surgical procedures move from the inpatient to the outpatient setting. The number of post-operative hospital visits most certainly decline if not disappear altogether for the typical patient. A change in the primary site-of-service may also affect the amount of pre-service work, but to a lesser degree than post-service work. A second objective is to identify services whose work effort may have changed as the speciality mix of physicians providing the service has changed, e.g., a shift from general surgery to interventional radiology for selected procedures. And, a third objective is to
identify services whose work has become less intense because of experience, i.e., capturing the learning-by-doing curve, or whose relevant patient population may have changed over time. The latter change could result through diffusion into a more broadly defined patient population or contraction to a more narrowly defined population. Services affected by these changes might be identifiable through an analysis of change in frequency of services over time. All three changes could lead to change in total work effort, pre- and intra-service work or time, or in the number and level of post-operative visits.

The seventh method described in Section 2.7 identifies services with potentially misvalued work RVUs through a direct comparison of work per unit time (WPUT). The objective of this method is the identification of statistical outliers — those services with a WPUT estimate that differs significantly from the typical value for a group of services expected to have similar levels of WPUT. This method builds upon the assumption that two services involving a similar technical approach, the same types of patients, and requiring approximately equal amounts of time should also be similar in physician work, or in other words these services should be similar in intensity. Further, the work of two clinically similar services requiring different amounts of time would likely vary proportionately with time — the work values would differ but work intensity would be similar. The Harvard researchers evaluated WPUT within narrow clinical families and across broad clinical families and concluded that directly rating service intensity is a difficult cognitive task and can lead to invalid estimates of intensity and work. The empirical findings from the original RBRVS research suggest that structured comparisons of WPUT may have value in assessing the appropriateness of work RVUs obtained from physicians' judgments of work. This is particularly true when these comparisons are made within groups of clinically similar services and for comparable periods of total service.

A common requirement across all seven proposed methods is the need to assemble databases describing the work and time estimates collected as part of the RBRVS research efforts over the past decade. We conclude this chapter by describing in Section 2.8 the need for a comprehensive database that includes not only baseline information, but also information gathered from any of the proposed
Methods for Identifying Misvalued Work Relative Value Units (RVUs)

methods that HCFA may undertake in support of its next 5-year review. We describe a number of different types of databases that HCFA might wish to consider developing, the primary level of information to be included, and the types of information that would be useful to HCFA going forward in their updating activities.
2.1 The Role of Clinical Panels in Evaluating Total Physician Work

2.1.1 Overview of the Method

Reevaluation of relative value units (RVUs) in the 5-year review inevitably involves clinical input. It could come in the form of the AMA’s Relative Value Update Committee (RUC) as it was done in the first 5-year review, in HCFA-formed panels that either are strictly internal, including Carrier Medical Directors, or external in the form of outside practitioners. This method reviews the various forms that clinical panels might take. The composition and deliberations of the panels may depend upon their role in the review cycle. Emphasis in this chapter is placed on HCFA panels that would play an initial role in identifying potentially misvalued CPT work values. Outliers would likely be sent to the RUC for detailed review. They may also be candidates for quantitative analysis at HCFA. HCFA also convenes panels after the RUC submits its recommendations. How HCFA might compose these panels is discussed later in Chapter 3.

This section of Chapter 2 concentrates on panel composition and mode of deliberations using the standard magnitude estimation approach. Section 2.2 introduces an alternative, simplified method for identifying misvalued procedures.

2.1.2 Background

Clinical panels have been used to set work RVUs for many years. In its development phases, Harvard and the AMA used many different types of panels in determining total work RVUs. The Harvard group even conducted a systematic study of the reliability of results across panel types (Leape et al., 1992).

Once the Medicare fee schedule (MFS) became operational, the RUC, which was formed in 1991 (AMA, 1998), began using a large multi-specialty panel to make recommendations to HCFA regarding work RVUs for new or revised codes based on revisions made to the CPT coding system by the CPT Editorial Panel. The RUC is composed of 28 members, 23 appointed by the national specialty societies, and 5 more allotted to the RUC Chair, the co-chair of HCPAC (an advisory
The Role of Clinical Panels in Evaluating Total Physician Work

committee representing nonphysicians and osteopaths), the AMA, the AOA, and the CPT Editorial Board.

Between January and August, 1992, a group from Rand convened four separate physician panels to pilot test a more streamlined, multi-specialty version of the Harvard panels (Kahan et al., 1994). Two panels focused specifically on urologic and ophthalmic procedures and were composed of a mix of procedure specialists and other surgeons. A third, multi-specialty panel had 13 members representing the major specialty societies while a fourth, pediatric panel had several pediatricians, surgeons, and primary care physicians. These panels were formed, in part, to study the dynamics of group deliberations. Their purpose was to guide HCFA in establishing its own panels.

Finally, HCFA, itself, has formed at least two kinds of panels, one of medical directors, and another with a mix of medical directors, central office physicians, and outside general and surgical specialists (sometimes referred to as refinement panels). These panels review RUC recommendations.

The panels that have been used to establish and/or evaluate work RVUs differed along several key dimensions highlighting issues in formulating new panels. Assembling a panel to collect subjective estimates of physician work requires decisions regarding:

- Panel composition (e.g., specialty mix, academic orientation, geographic representation, gender, age);
- Panel size;
- Nature of member interaction (e.g., face-to-face, telephone survey, mail survey, mail plus face-to-face);
- Method used to gain consensus, if any; and
- Method for rating work.

**Survey versus Face-to-face.** Large numbers of physicians have been surveyed both by mail and telephone to collect impressions about relative work of preselected codes. The RUC continues to ask medical societies to survey its members by mail to collect data on times, complexity, and total work in order to justify recommendations for RVU changes. (We are not aware of HCFA actually
conducting its own surveys to collect information on physician work.) Surveys can be further distinguished by the number of respondents and the extent of resurveying. Clinicians can be surveyed by mail or telephone once, twice, or even three times. If more than once, the surveys are usually carried out in a Delphi approach by returning the results of the whole set of respondents to each participant and asking for another rating—again anonymously. Surveys could be administered to 100 or more physicians or to as few as 10 (Leape, et al., p. ns35). Conversely, panels of physicians could be assembled for face-to-face rating of codes with no prior, broader survey of practitioners. Finally, a combination of survey and small-group face-to-face approaches can be used.

**Delphi versus Consensus.** In a strict Delphi interview format, panelists are given statistical feedback on group ratings and rerate work in a second, and possibly even third round without face-to-face contact or verbal discussion. A modified Delphi approach brings together raters for a final face-to-face discussion of the ratings previously completed anonymously by the panel (Kahan et al., 1994). A consensus panel takes the Delphi technique a step further and encourages the panel to reach a final consensus, or unified agreement, on the ratings of each code. (Papal conclaves are probably the best-known examples of consensus panels.)

**Magnitude Estimation.** In previous Harvard studies, panelists’ ratings of total work have used magnitude estimation techniques that ask: “Compared to [a reference procedure] rated at 100, how much more or less work does procedure X involve?” The respondent would report that a procedure might be a 90, or a 150, i.e., 10 percent less or 50 percent more, on an open-ended scale. Absolute scores are converted to RVUs by multiplying by the actual RVUs of the reference procedure and dividing by 100.

The RUC’s current survey format uses a very similar approach except panelists work strictly in RVUs (AMA RVS Update Survey, 1999). The respondent is given a list of current total work RVUs for a preselected set of reference services. He/she is then asked to estimate the amount of pre-, intra-, and post-service times for the new code and a reference set code. The rater is also asked to

---

The Identification of Potentially Misvalued Work RVUs: 2-9

Health Economics Research, Inc.

The Rand panels were designed to test, explicitly, changes in preliminary ratings made independently by members before face-to-face discussions.
report the number of visits post-operatively for global surgical codes and to rate the complexity of each of the three phases of the service on a 1-to-5 Likert scale of complexity. Finally, the rater is asked to give an estimate of the total work RVUs of the new/revised code, presumably based on the relative time and complexity data reported earlier in the survey. No explicit link is required, however, between the time and complexity ratings and the final, global total work RVUs.

The four Rand panels also worked directly with work RVUs. Members were given a visual "ruler" that scaled 10-20 reference set procedures according to total work. Panelists were to fit additional codes on the ruler.

Some debate has taken place over the advantages of an "inclusive" estimate of total work versus a "building block" approach. Harvard's early research asked raters to rate the time and intensity of the three phases separately. Time was multiplied by relative intensity to produce "time-intensive" minutes that were later scaled to a numeraire procedure to produce RVUs. Validation studies found that respondents were double counting time to some extent by incorporating it into intensity. That is, when asked to rate a procedure simply by total work versus time and intensity separately, the inclusive approach generally produced less effort on the physician's part. In its later phases, the Harvard researchers switched to estimates of total work. The current AMA survey follows the same strategy, essentially forcing the rater to incorporate both the time and complexity into a single number relative to the total work RVUs provided for the reference procedure. Separate reports of procedure time and complexity presumably help guide the final RVU rating by the respondent.

**Panel Validity and Reliability.** Leape and his colleagues (Leape et al., 1992) conducted a study of various Delphi and consensus panel approaches using a large national telephone survey of many specialties as the gold standard. What they found was somewhat surprising.

*Single-round mail survey of a small number of experts provides relative work values that compare more favorably to those of the national survey than those obtained from Delphi multiple-round ratings or modified Delphi with face-to-face discussion.*
The single-round mail survey produced results that disagreed roughly 11 percent on average from nationally collected total work RVUs. When a second panel was given its initial ratings and asked to rerate based on first-round averages, the ratings began to deviate further from the national results. When the panel finally met to discuss their ratings and rerate once again, the results were fully 21 percent different from the national results (Leape et al., 1992, Table 4). Finally, when the panel was asked to reach a consensus on the ratings, they were able to do so on all of them, but the consensus ratings produced nearly a 24 percent deviation from national figures. The more the panelists talked face-to-face the worse the ratings became—relative to the gold standard.

Leape et al. (1992) also found: (1) ratings for higher valued services tended to increase with each iteration while lower valued services decreased; (2) the intra-class correlation in the panels was quite high; and (3) the ratings of small panels of 11 or 19 surgeons differed insignificantly from those based on a hundred physicians in the national survey. These results supported the authors' conclusion that mailed surveys filled out anonymously and independently by a very small group of well-chosen practitioners is more consistent with results from a large random survey than any results obtained through Delphi or consensus panels—especially meeting face-to-face. They also concluded that a single-round mailed survey is highly cost effective.

Leape et al. (1992) explain their findings on poorer face-to-face performance based on personal observation buttressed by the psychological literature on small-group decision making. First, Delphi numerical feedback exerts more conformity, possibly focusing on the wrong answers (p. Ns36). Members who are highly articulate with dominant personalities can bias panelist ratings. Experts often feel the need to defend their high ratings for more complex tasks, especially when higher ratings have direct impact on fees and incomes of the profession. Panelists who rarely perform highly weighted procedures have little to lose by upcoding their ratings based on the discussion.

Another key finding by Leape and his colleagues is that variation in the small group panel ratings after the first round mail surveys was less than in the large national survey. This they attributed to the careful selection of participants. Nominations for the study were solicited from the
The Role of Clinical Panels in Evaluating Total Physician Work

six major regional surgical societies. Both academic and community-based general surgeons who were board certified and highly respected in their communities were recommended. From 60 nominees, two panels of 19 and 11 surgeons were randomly selected to be geographically balanced with a mix of academic and nonacademic members.

The Rand researchers came to a somewhat different conclusion than Leape’s group; namely,

A panel process for refining [work RVUs] is practical, provided that panelists are provided with a valid reference set for comparison purposes and provided that all members feel comfortable engaging in the discussion. (Kahan et al., 1994, p.1082)

The Rand researchers admittedly experienced considerable difficulty with the ophthalmology panel where the ophthalmologist “performers” of the procedures dominated the discussion and “rej ected cross-specialty comparisons of work.” (Kahan et al., 1994, p.1083) It became clear in this panel how important it is to have panel members who accept the reference scale and who are willing to challenge specialists beyond their own areas of expertise (Kahan, et al., 1994, p.1083).

Clearly, a high priority for the establishment of a flexible and responsive RBRVS revision procedure is the existence of a statistically validated, consensually accepted reference set. (Kahan, et al., 1994, p.1083)

At present, the usual approach to conducting the annual and 5-year reviews has been to form face-to-face panels to “iron out” if not reach total consensus, on RVUs for new/revised or misaligned codes. The RUC is a consensus panel of sorts in requiring a two-thirds vote to approve a new/revised value for total work to be sent on to HCFA. HCFA, in turn, generally uses a similar consensus approach by convening “refinement” panels of government and outside clinical experts. Interestingly, HCFA’s acceptance rate for RUC recommendations has increased to over 90 percent, which is presumably due to a more rigorous, uniform approach being used by the RUC. How much is also due to the natural “drift” of face-to-face panels towards a personality/expert-driven result is unknown.
The Role of Clinical Panels in Evaluating Total Physician Work

The immediate purpose of panels in the first phase of the 5-year review differs in simply identifying, rather than correcting, potentially misvalued procedures. A notable exception might be in a recalibration of the reference set work values prior to RUC deliberations on other codes. Phase I panels may be necessary to realign reference code RVUs.

2.1.3 Data Requirements

The proposed method(s) uses ratings obtained from physicians (and possibly other clinicians) through either a mail survey or a face-to-face meeting to both identify misvalued codes and obtain estimates of new work RVUs. To the extent quantitative data on operating room times, number of post-surgical visits, etc. are available, this information could either be provided to the panel or used to evaluate the reasonableness of the panel’s ratings.

2.1.4 Detailed Description of Method

In the next 5-year review, clinical panels will likely be formed to identify rank order anomalies within and across clinical families. They may also be formed to investigate the accuracy of RVUs for codes that are performed by more than one specialty. Finally, panels may also be used to reevaluate the reference set, or multiple points of comparison, that underlie the entire relative value system.

Below we provide step-by-step instructions for conducting both mail surveys and face-to-face meetings to identify misvalued services. We describe identifying within-family rank order anomalies using a single-round mail survey method. Two alternative rating methods are suggested for identifying potentially misvalued work codes: magnitude estimation, and Rasch paired comparison. The latter method is discussed in detail in Section 2.2 below. Magnitude estimation is described in this section. A face-to-face method is recommended for identifying and correcting misvalued services that are provided by multiple specialties. Multi-specialty panels are also recommended for recalibrating the reference set.
The Role of Clinical Panels in Evaluating Total Physician Work

Within-family Rank Order Anomalies

Step 1: Identify the subsets of codes that will be examined by a given panel.

Step 2: Identify the specialties that have some familiarity with the work involved in most, if not all, of the family of codes. Surgical codes should be examined by panels of general surgeons or specialists depending upon the family of codes.

Step 3: Ask the relevant national speciality societies for nominations of experts to participate in one of several different panels. In requesting nominations, ask for representatives of both academic and community-based medicine and surgery. A broad mix of nominees should be requested that practice in each of the nine census divisions and in urban and rural areas. Not only does a broad mix maximize the representativeness of the resulting ratings, it also is more defensible and less open to criticism from the medical community.

Step 4: Select a random number of physicians for each panel. Fifteen to twenty are adequate, assuming no nonparticipants. With rejections, replace with random sampling to achieve at least 15 participants.

Step 5: Mail the survey instrument to the panelists closely following the RUC survey method in asking for the pre-, intra-, and post-service times, complexity, and total work based on pre-selected reference procedures most relevant to the family of codes in question. Respondents should understand that the time and complexity estimates are only a guide in determining the final total work RVUs using reference set RVUs as ultimate benchmarks. This is to avoid double-counting of time and complexity.

Step 6: Collect and process the responses from the panels and compare the mean RVUs to those published in the most recent Federal Register.

Step 7: Identify “outlier” codes in terms of work (e.g., those deviating by more than 15-20 percent).

Step 8: Using relevant quantitative data collected as part of the 5-year review, analyze outlier codes for changes in time and complexity.

Step 9: Based on quantitative analysis, either ignore an outlier or send it on to the RUC for further review. Send to the RUC only codes where quantitative data support a possible misvaluation.
The Role of Clinical Panels in Evaluating Total Physician Work

Cross-specialty Anomalies in Similar Services

Step 1: Based on the same national specialty recommendations used in the within-family analysis, convene a panel of general surgeons and related specialties (e.g., cardiology, radiology).

Step 2: Ask the panelists to identify sets of services that should have very similar total work ratings. The panel would be guided by commonalities among the services (e.g., endoscopies, imaging, anesthetic injections, laser surgery). Essentially identical procedures done on adults and children could also be identified for comparison.

Step 3: Conduct a face-to-face study of pairs of supposedly similar codes. Ask each panelist to rate codes in terms of total work using magnitude estimation (or Rasch, see below) techniques. Do not seek consensus.

Step 4: Give the panel actual RUC or Harvard times plus the current and predicted RVUs for the codes being compared and ask them to explain differences in the procedures that require more or less work.

Step 5: Send list of potentially misvalued codes across specialties to the RUC for intensive review.

Recalibrating the Reference Set Procedures

One criterion in choosing a reference set procedure was stability; procedures should not be experiencing significant technical or organizational change. Yet, over time, the RVUs for some of the reference set codes can change—possibly due to technical change or substitution of non-physician personnel. This is especially true of some of the more technically complicated procedures such as CABG where lengths of stay have declined rapidly and non-physicians are performing more tasks in the operating room. Re-validating the reference set ought to occur prior to the review of the rest of the MFS work values. This is because RUC review of other potentially misvalued codes keys off of the reference set. Examining the reference set also requires a broader set of specialties than for any family of codes, although not to develop cross-specialty links as before.

---

3 The RUC could review non-reference set codes without recourse to the reference set (see Section 3.2.1 below) but likely would prefer to continue to anchor surveyed physicians by providing revised reference set RVUs.
The Role of Clinical Panels in Evaluating Total Physician Work

Harvard-AMA convened several panels of different specialists to, first, identify cross-specialty links and then to adjust for specialty-specific valuations of some of the reference procedures in terms of work (Braun et al., 1988). Panelists identified services across specialties that were equal (same service, different specialist) or equivalent (different service, same work) in terms of work. These procedures formed links between the rating systems of the separate specialist panels that had previously rated codes within narrow families. Eventually, 275 links were identified, forming 550 linked pairs. The separate specialty ratings of intra-service work for each of these 275 link procedures were used in a regression to determine inter-specialty work multipliers. Multipliers were needed because the specialty panels used different reference procedures in scaling their codes. As a result, general surgeons, for example, may have rated their linked procedures 25 percent more work, on average, than, say, OBGYNs because the latter rated work based on a more difficult procedure than general surgeons. All specialty ratings were then placed on a common single scale by adjusting specialty-specific ratings of work by the link multipliers.

Rand researchers, under HCFA contract, developed a modified linking process that was somewhat simpler (Morton, Kominski, and Kahan, 1994). They worked directly with CPT codes rather than first having to convert vignettes to codes. They also required that the links be transitive. If A involves more work than B which involves more work than C, then A must also involve more work than C. This was not always the case in Harvard’s work. Linked codes must have the same RVUs. And unlinked specialty services should have similar relative work on the common scale compared to the original specialty relative ratings. Rand dropped intensity-based definitions of similar work that were occasionally used by Harvard.

Rand and Harvard linking methods differed fundamentally in what could vary on the common scale. Rand forced all linked procedures to have the same work values and allowed the within-specialty relative ratings to change. This had pronounced effects on certain specialties such

---

4 Harvard identified links using three different definitions of work: (1) total; (2) intra-service; and (3) intensity.
as anesthesia whose work values became quite overvalued. Harvard forced the within-specialty
relative ratings to remain unchanged and shifted the work values of equivalent codes.

Re-examining the reference set using clinical panels could involve the following steps.

Step 1: Solicit public comments on the reference set as soon as possible. Limit comments only
to changes in work effort only for specific codes. Prohibit comments that whole families
are misaligned relative to other specialty codes.

Step 2: Request any additions or deletions to the reference set from the RUC. Limit RUC direct
input to suggested changes in the list and not any work RVUs.

Step 3: Simultaneous with soliciting public comments, conduct quantitative analyses of all
reference set codes (see Sections 2.3-2.7). Identify reference set codes that may have
undergone changes in work effort.

Step 4: Form 3 panels of physicians to review various components of the reference set for recent
misalignments. The panels should represent all of the major specialties, with
approximately 12-15 members per panel. Panel composition should be staggered similar
to the original Harvard multi-specialty panels with primary care and Medicare Carrier
Medical Directors on all panels with other panelists drawn from various medical and
surgical specialties.

Step 5: Distribute materials in advance of the panel meetings. Materials would include (a)
summaries of the quantitative findings on changes in lengths of stay, visits, OR times, etc.,
and (b) a set of guidelines for evaluating the 6 factors underlying changes in work:
technology diffusion, learning-by-doing, technology substitution, personnel substitution,
re-engineering the service, and patient severity (see Section 3.2.2 for more details).

Step 6: Convene face-to-face panels to review hypothesized changes in work effort. Identify
panelist to lead discussion of potential changes for given family of codes, e.g., heart
surgeon, ophthalmologist. HCFA-selected moderator should assure all six underlying
factors affecting work have been covered in the discussion.

Step 7: After discussion of family of codes, conduct independent, secret vote on percent change
in total work in each code separately. Many codes will likely have no change.

Step 8: After all codes have been reviewed and scored in a panel, allow panelists to revisit all
their percent change scores to assure fairness across families of codes.

Step 9: Calculate average percent change in total work for each reference set code. Use Tukey
bi-weights in calculating mean or use median value.

Step 10: Adjust all reference set RVUs by panel average (median) recommended percent change.
2.1.5 Illustrative Example

Small panels similar to those used by AMA-Harvard in Phase III of their original construction of work values can use magnitude estimation to identify potentially misvalued procedures. Table 2-1 provides an illustrative example of how 15 panelists might rate a family of 12 hand-and-finger incision codes. An outpatient office visit with an established patient for 15 minutes is included as well for benchmarking purposes (explained in a moment).

First, each panelist is given a listing of the codes and CPT numbers. The list includes three starred codes (in addition to the office visit) that also appear on the reference set. The pre-established average ratings of work for these reference codes (in bold) are to be used by the panelists as anchors for rating the remaining hand/finger incision procedures.

Next, the 15 panelists would each rate the work of each code relative to all three reference set codes. Their coding would be open ended with no upper limit. Panelist 1, for example, hypothetically rated code 26060, single percutaneous Tenotomy, as slightly more work than code 26040, percutaneous palmar Faciotomy (600 v. 562), while panelist 2 rated the former as slightly less work. Once the panelists have completed their raw coding, the data are logged (to base 10) to better reflect a normal distribution. Outlier ratings can be handled either by dropping those outside, say, 3 standard deviations, or weighted using bi-squared weights that give less weight to outlier responses. With small panels, retaining outlier values is desirable. A mean (logged) rating is calculated using the weights along with an approximate standard error.

Next, a geometric mean rating is calculated by exponentiating (from the base 10) the mean logged value. The result is shown in the Geometric Mean Rating column in the table. In the illustrative example, the physicians in the panel rated code 26060, Tenotomy, single percutaneous, almost identical to reference set code 26040, Faciotomy, palmar, percutaneous. Code 26010, Drainage of finger abscess, is rated lowest in the family (246), or slightly more than half as much work as reference code 26055, Tendon sheath incision.

---

1 See Hsiao et al. (1993) for a brief discussion of the formulas used to derive the weights. The weights are a nonlinear transformation of a deviation statistic equal to the absolute difference between the mean and an outlier rating, divided by a modified interquartile range.
### The Role of Clinical Panels in Evaluating Total Physician Work

#### Table 2-1

**Illustrative Rating of Hand Procedures' Total Work by 15 Panelists with Conversion to RVUs**

<table>
<thead>
<tr>
<th>Procedure/Service</th>
<th>Panels</th>
<th>Geometric Mean Rating</th>
<th>RVUs(^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td><strong>99213</strong> Office visit, est. patient</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>26010 Drainage of finger abscess, simple</td>
<td>200</td>
<td>300</td>
<td>250</td>
</tr>
<tr>
<td>26011 Drainage of finger abscess, complic.</td>
<td>300</td>
<td>400</td>
<td>400</td>
</tr>
<tr>
<td>26034 Deep incision, hand/finger</td>
<td>1330</td>
<td>1150</td>
<td>1200</td>
</tr>
<tr>
<td>26035 Decompression fingers</td>
<td>1600</td>
<td>1500</td>
<td>1450</td>
</tr>
<tr>
<td>26037 Decompressive fasciotomy, hand</td>
<td>1550</td>
<td>1400</td>
<td>1300</td>
</tr>
<tr>
<td>26040 Faciotomy, palmar; percutaneous</td>
<td>*</td>
<td>*</td>
<td>***</td>
</tr>
<tr>
<td>26045 Faciotomy, open; partial</td>
<td>*</td>
<td>*</td>
<td>***</td>
</tr>
<tr>
<td>26055 Tendon sheath incision</td>
<td>*</td>
<td>*</td>
<td>***</td>
</tr>
<tr>
<td>26060 Tenotomy, percutaneous, single</td>
<td>600</td>
<td>500</td>
<td>575</td>
</tr>
<tr>
<td>26070 Arthrotyomy, exploration, drainage</td>
<td>800</td>
<td>700</td>
<td>850</td>
</tr>
<tr>
<td>26075 Arthrotyomy, metacarpophalangeal joint</td>
<td>850</td>
<td>800</td>
<td>900</td>
</tr>
<tr>
<td>26080 Arthrotyomy, interphalangeal joint</td>
<td>900</td>
<td>950</td>
<td>900</td>
</tr>
</tbody>
</table>

**NOTES**

\(^a\) RVUs based on 0.55 * Geometric Mean Rating / 100.

\(^b\) Actual RVUs for reference set procedures.

* Reference set procedures.
Finally, the open-ended scale must be converted to RVUs. This is done by multiplying the RVUs for code 99213, an established patient 15-minute office visit, equal to 0.55, by the geometric mean rating divided by 100. The bolded numbers for the three family-specific reference procedures (3.09, 5.27, 2.56) come from the 1995 RVUs used in the first 5-year review. The rest of the RVUs are illustrative and do not necessarily conform to the actual RVUs now being used for these codes.

The ultimate RVUs produced by the panel can be used either to replace the current RVUs in the MFS or to identify potentially misaligned work values. If the latter were the short-run goal, then the RVUs would simply be compared with the current RVUs and outliers identified. For example, the highest rated code outside the reference set was code 26035, Decompression of fingers (8.33). In the current 1997 MFS this code was rated at 9.51 RVUs. Thus, 26035 would appear overvalued and would be a candidate for special investigation by the RUC or other HCFA staff.

2.1.6 Strengths and Weaknesses of Method

Strengths

- The single-round survey method with a limited number of 15 respondents has been validated (Leape et al in Medical Care 1992, issue edited by P. Braun) in a peer-reviewed publication, thus giving this method face validity.

- The method of single round self-administered questionnaires (services rated vs. set of benchmarks in the specialty) was used extensively in Phase 3 of Harvard's RBRVS study.

- Mailed surveys are far less costly than convening many meetings of clinicians for one or two days.

- One-round mailed surveys protect against undue influence of specialists.

Weaknesses

- Subsequent rounds of surveys and face-to-face methods (i.e., a modified Delphi technique) result in divergence from "gold standard" values.

- Except for reference set, the limited size of the suggested panels should not be considered definitive enough to immediately recalibrate misaligned RVUs. Identifying misaligned RVUs for further RUC review seems more appropriate.
The Role of Clinical Panels in Evaluating Total Physician Work

- It is assumed that the RUC use of quasi-consensus panel formats is not subject to change. Any RUC recommendations after reviewing potentially misaligned procedures may reflect biases inherent in consensus panels.
- Magnitude estimation is more challenging and time consuming for panelists than paired comparison methods (discussed in the next section).

2.1.7 Likely Response By Key Stakeholders
- Specialists may have a preference for face-to-face meetings (modified Delphi) as is done in the RUC process, even though this method has been shown to introduce divergences and biases toward dominant discussants or persuasive advocates of particular procedures.

2.1.8 Time Frame

There are a number of data collection modes that may be used to obtain estimates of physician work from clinical panels. We are recommending a single round mail survey of fifteen clinicians for identifying rank order anomalies in physician work within small clinical families, and a face-to-face panel meeting of fifteen clinicians for identifying cross-specialty anomalies in similar services. Either data collection mode could be used to evaluate the reasonableness of current work RVUs for the multi-specialty reference set. However, we believe a face-to-face meeting of a multi-specialty panel would be a superior data collection mode for identifying problems within the reference set and for developing new work RVUs for these services. Because the timing of specific tasks differ between the two proposed data collection modes, we describe them separately below.

It is anticipated that identification of inappropriately valued work RVUs using a single round mail survey of fifteen clinicians can be completed within four months from the start of the project. Exhibits 2-1 and 2-2 display the timing of the tasks described in Section 2.1.4 for conducting a single round mail survey and a small group meeting, respectively.
### Exhibit 2-1

**Timeframe for Conducting a Single Round Mail Survey of Clinicians**

<table>
<thead>
<tr>
<th>TASK</th>
<th>TASK DESCRIPTION</th>
<th>Month 1</th>
<th>Month 2</th>
<th>Month 3</th>
<th>Month 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Identify the Codes to be Examined by the Panel</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2-4</td>
<td>Assemble a Small Panel of Clinical Experts</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Develop and Distribute Information to the Panelists</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Collect and Process the Responses from the Panelists</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7-8</td>
<td>Identify Outlier Codes from Panelists and compare with</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Quantitative Data Collected from Other Proposed Methods</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Send to the RUC for structured review codes</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**KEY**

- ▲ Completion of Task
<table>
<thead>
<tr>
<th>Task</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Develop Sets of Services that should have similar total work</td>
</tr>
<tr>
<td>2</td>
<td>Conduct a Face-to-Face Study of Pairs of Ratings of Similar Codes</td>
</tr>
<tr>
<td>3</td>
<td>Send to the RUC for structured review</td>
</tr>
</tbody>
</table>

**Exhibit 2-2**

Timeframe for Conducting a Small Group Face-to-Face Meeting of Clinicians

<table>
<thead>
<tr>
<th>Month 1</th>
<th>Month 2</th>
<th>Month 3</th>
<th>Month 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>14</td>
<td>15</td>
<td>16</td>
</tr>
</tbody>
</table>

**KEY**

△ Completion of Task
Clinical Panels using a Single Round Mail Survey

Task 1: Identify the codes to be examined by the panel

We estimate that it will take approximately one month for HCFA to decide which codes to be sent to clinical panels for structured review. During this time, HCFA must identify the specific codes to be evaluated, identify appropriate reference set codes, determine the number of clinical panels that will be required, and identify the necessary specialities to populate the panels. It is anticipated that a number of different methods would be used by HCFA to identify potentially misvalued services (see sections 2.2. through 2.7). This one month time period would be spent comparing potentially misvalued services identified across all of the methods and assembling a list of unduplicated services for clinical review.

Tasks 2-4: Assemble a small panel of clinical experts

Once the number and composition of the panels have been determined, we estimate that it will take approximately two months to assemble the panels of clinical experts. During this time, we recommend HCFA seek nominations from a variety of different physician organizations, e.g., the American Medical Associations Relative Value Update Committee, specialities societies, and local medical associations, to ensure broad representation to the panels. Materials need to be distributed to the targeted physician organizations specifying the qualifications of the physicians being sought, the timing of the necessary tasks, the likely level of effort, etc. Follow-up with nominated physicians to ensure their interest and availability also will be necessary as will written confirmation of their appointment to a panel, once the final selections have been made.

Task 5: Develop and distribute information to the panelists

We estimate that the mail survey can be conducted during the third month. Development of the mail survey instrument could begin at the end of Task 1 and be completed during the first week of this task. At the beginning of the second week, we recommend sending the mail survey instrument to all panelists via Federal Express to ensure timely delivery and highlight its arrival.
The Role of Clinical Panels in Evaluating Total Physician Work

within the physician's office or at their home. Panelists should be requested to complete and return the survey within a week and a Federal Express return envelope provided for their use. By the end of the third week, telephone reminders should be made to panelists who have not returned the survey instrument. These panelists should be requested to complete and return the survey instrument by the end of the fourth week. Failure to do so will result in their responses being excluded from the subsequent analysis. The fourth week also provides an opportunity for a remailing, if any panelists have misplaced their initial survey instrument.

Task 6: Collect and process the responses from the panelists

It is anticipated that collecting and processing the responses from the panelists should be completed within two weeks of the survey ending. Development of tabular presentation materials of the results should occur during these two weeks, displaying the mean values obtained from the panelists and current work RVUs. To the extent that objective data were provided to the panelists or have become available to aid in the evaluation of the panelists' responses, they should be incorporated into the presentation materials.

Tasks 7-8: Identify outlier codes from panelists and compare with quantitative data collected from other proposed methods

It is anticipated that an additional two weeks would be spent identifying outlier codes from the ratings obtained from the panelists and comparing the panelists' estimates of physician work with underlying objective data, if available.

Task 9: Send to the RUC for structured review

The last step in this method is the submission of identified services to the RUC for their review and HCFA review of any proposed new work RVUs. A specific time frame for the RUC deliberations and HCFA review is not included in Exhibit 2-1 as it can very depending upon the number of services submitted to the RUC, the number of specialties that need to be surveyed, and
availability of RUC internal resources. In Chapter 3, we provide a more general discussion of the timing of these activities within the context of the next 5-year review.

*Clinical Panels using a face-to-face panel meeting*

**Task 1: Assemble a small panel of clinical experts**

The first step will be to determine the number and composition of the panels that will be used in this method. Similar to the identification of small panels for the mail survey collection mode, we estimate that it will take approximately two months to assemble the necessary panels of clinical experts. During this time, we recommend HCFA seek nominations from a variety of different physician organizations, e.g., the American Medical Associations Relative Value Update Committee, specialities societies, and local medical associations, to ensure broad representation to the panels. Materials need to be distributed to the targeted physician organizations specifying the qualifications of the physicians being sought, the proposed dates for the meeting, and the location of the meeting site. Follow-up with nominated physicians to ensure their interest and availability also will be necessary as will written confirmation of their appointment to a panel, once the final selections have been made.

**Tasks 2-4: Develop sets of services that should have similar total work ratings and conduct a face-to-face study of pairs of ratings**

We estimate that tasks 2 through 4 will be completed in one day face-to-face meetings. It is unlikely that all panel meetings can be scheduled to occur simultaneously. Therefore, a two month window for the panel meetings is recommended. It is anticipated that processing the responses from the panelists should be completed within one week of the panel meeting. Development of tabular presentation materials of the results should be completed within another two weeks following the panel meeting.
Task 5: Send to the RUC for structured review

The last step in this method is the submission of identified services to the RUC for their review and HCFA review of any proposed new work RVUs. A specific time frame for the RUC deliberations and HCFA review is not included in Exhibit 2-2 as it can very depending upon the number of services submitted to the RUC, the number of specialties that need to be surveyed, and availability of RUC internal resources. In Chapter 3, we provide a more general discussion of the timing of these activities within the context of the next five year review.

2.1.9 References


2.2 A Rasch Paired Comparison of Total Physician Work

2.2.1 Overview of the Method

This method identifies misvalued CPT codes in terms of either total or intra-service work based on a small-group panel comparison of codes within families. The objective is to identify statistical outliers that appear to be either misaligned or compressed in terms of overall physician work effort employing a simpler approach than magnitude estimation. The method is based on original research undertaken by Dr. Robert Florin, a retired neurosurgeon and current member of the AMA RUC. Rasch paired comparison combines methods developed from educational research and statistical regression to, first, reorder families of codes, then convert the new cardinal scoring system to total work RVUs. This method is likely to be only one of a number of psychometric methods that could be used to obtain estimates on physician work. Time and resource constraints precluded a more exhaustive search of alternative methods.

2.2.2 Background

After the first 5-year review of the Medicare fee schedule, Dr. Florin began a systematic study of new ways of identifying ranking anomalies among families of codes as well as responding to the general impression among many surgeons that the RBRVS scales were compressed in certain families—and possibly across families and specialties. The RUC approach to generating RVU weights for new procedures suggested a form of paired comparison. That is, the RUC asks surveyed physicians (usually through medical societies) to report the estimated time and complexity bounded by the two reference set procedures closest to the new code in question. Recognizing the burden on respondents and problems of sample bias, Dr. Florin turned to an alternative psychometric method for generating a cardinal rank ordering of small families of procedures. This procedure does not require extensive surveys; only a modest number of participants (20-30) who could quickly fill out a page or two of comparisons as part of another meeting.

The underlying approach is called Rasch measurement analysis, first developed by Georg Rasch in 1950s-1970s. It has been used in a wide variety of disciplines including education and test
A Rasch Paired Comparison of Total Physician Work

grading, health outcomes research, physiology, psychophysics, writing performance, mathematics, marketing, physics, and ethical valuation. Paired comparison is a sub-analysis under the broader rubric of Rasch analysis. Rasch methods are grounded in the psychometric literature dealing with how to “count things.” Rasch measurement models are based on a simple performance measure, \( L_n = B_n / D_i \), where \( B_n \) = the level of ability of the n-th person (i.e., test taker or rater) and \( D_i \) = the difficulty of the i-th test item (Andrich, 1988). The measure incorporates both the abilities of the test taker (if relevant) and the difficulty of the items (questions) on the test. The greater the ability of the test taker, the higher the Rasch performance measure while the greater the item difficulty the lower the score or rating.\(^6\)

The probability of observing a particular Rasch score is assumed to be distributed logistically as \( \varphi(L_n) = L_n/(1 + L_n) = (B_n/D_i)/G_n \), where \( G_n = 1 + (B_n/D_i) \) is a normalizing factor. On any item scored zero or one (e.g., right versus wrong, more versus less), the probability of scoring a 1 is \( L_n/(1 + L_n) \), while the probability of a zero is \( 1/(1 + L_n) \). Note that the two probabilities sum to 1.0. Thus, the Rasch is a cumulative model. As the tester’s ability increases relative to the difficulty of the test items, the measure increases. Conversely, the more difficult the test items relative to a tester, the lower the score.

Rasch measurement is flexible enough to rate both the abilities of different testers and the relative difficulties of the test items. In evaluating physician work, we would like differences in raters’ “abilities” (i.e., perceptions of procedure difficulty) to cancel out, leaving just the relative difficulties of the procedures. This can be done by calculating the odds of a person’s rating of work on two items:

\[
\begin{align*}
\text{Pb}(1,0) &= (B_n/D_i)/G_n = D_2/D_1 \\
\text{Pb}(0,1) &= (B_n/D_2)/G_n
\end{align*}
\]

where \( \text{Pb}(1,0) \) = the probability that a single rater would rate the first item a 1 (or more work) versus a second item (=0). Note that the rater’s ability cancels out when odds ratios are taken. Also note that the odds of rating item 1 as more work depends positively on the difficulty of item 2 and

\(^6\) For physicians rating the work involved between two procedures, the basic Rasch ratio is reversed, with higher difficulty a positive factor in the numerator and the physician’s ability to perform difficult procedures in the denominator.
A Rasch Paired Comparison of Total Physician Work

inversely with item 1. This is backwards and is easily corrected by recoding the procedure with more work as a zero.

Next, in rating two items, the likelihood of either being rated more work can be expressed as:

\[
\begin{align*}
\pi_{12}(1,0) &= \frac{1}{D_1/(1/D_1 + 1/D_2)} = D_1/(D_1 + D_2) \\
\pi_{21}(0,1) &= \frac{1}{D_2/(1/D_1 + 1/D_2)} = D_2/(D_1 + D_2)
\end{align*}
\]

where \(\pi_{12}(1,0)\), for example, is the probability of rating item 1 more work than item 2, conditional on the fact that the single rater is comparing only the two procedures and not both to a third procedure. Note that the rater’s ability cancels out in the numerator and denominator (leaving only 1’s). Multiplying through the top and bottom of (1a) by D2 and (1b) by D1 gives the last expression in the probability formulas. Thus, the probability of rating procedure 1 as more work than procedure 2 depends on the difficulty of procedure 2 as a proportion of the combined difficulty of the two procedures. (Again, the ratings need to be reversed if procedure 1 is regarded as more work.) Note that the two probabilities sum to 1.0.

To build up an estimate of the relative difficulties of the two items, Rasch relies on multiple respondents scoring the two items as more or less work (which will be expanded to include more items below). The probability of observing a given frequency of 0,1 responses for the two items can be determined using a binomial distribution, i.e.,

\[
P(b; f_{12}, f_{21}; D_1, D_2) = \frac{f_{12}!}{f_{12}!f_{21}!}\pi_{12}^{f_{12}}\pi_{21}^{f_{21}}(1-\pi_{12})^{f_{21}}
\]

where \(f_{12}, f_{21}\) = positive ratings for item 1 over 2 and 2 over 1, respectively; and \(f_{12}\) = the total number of times items 1 and 2 are rated. Eq. (2) gives the probability of observing the exact combination of positive ratings for items 1 and 2 (= \(f_{12}, f_{21}\)) based on the true underlying probabilities of observing the individual patterns, (1,0) and (0,1), for the two items. Note that the relative difficulty of the two items, or work effort, is embedded in the \(\pi\)-probabilities (which add to 1.0).

Eq. (2) can be expanded to consider all pairs in a set of many items, thereby producing a likelihood function of all pairs. The question then becomes: What is the set of individual
probabilities, $\pi_{ij}$, that maximize the likelihood of observing the set of rater 0,1 scores, $f_{ij}$, for all the individual pairs? This is determined by first taking logs of the overall likelihood function (derived by multiplying (2) by all ij combinations), differentiating with respect to the $\pi_{ij}$'s (or the $\delta_{ij}$'s = logs of the $D_{ij}$'s embedded in the $\pi_{ij}$'s), and setting the resulting equations for the i-items equal to zero:

$$-\sum_j f_{ij} + \sum_j F_{ij} \pi_{ij} = 0, \quad i=1,...,L.$$  

The constraints are then solved iteratively for the optimizing $\pi_{ij}$'s. In order to produce absolute ratings, instead of relative difficulties, a further constraint is imposed that the sum of the (logged) difficulties (the $\delta_{ij}$'s) are set equal to zero. Hence, some items will be scored negatively, in logs, and others, positively. Taking antilogs gives positive final item scores greater or less than 1.0. Because the probabilities are not independent, they must be solved iteratively to meet the conditions for the L equations implied in (3). The maximizing conditions implied by (3) require that the weighted sum of probabilities that a particular item will be preferred over other items (i.e., rated higher work) be equal to the total number of times the item is rated positively, or greater work, across all paired comparisons. The weights are the number of times a particular pair is evaluated (i.e., the $F_{ij}$). The $F$-weights effectively normalize for the number of times a particular item is rated, which is necessary in that an item will likely receive more positive scores the more times it is compared with other items. Relative, and absolute, difficulty of each procedure compared to others is determined by the frequency with which it is rated more difficult with respect to each of the other alternatives.

An added flexibility of the Rasch method is its linking ability when raters do not rate all pairs. Linking is crucial in devising tests of varying difficulty when different groups of test takers are not given all questions on a test (or may not respond to all of them because of their difficulty). This is called tailored testing. What is required is some overlapping items. So long as some items are rated by both groups (e.g., general and neurosurgeons), Rasch methods can extend the results to produce a single common scale.

Dr. Florin has personally conducted several small-group rating sessions using Rasch methods. In early 1998, he took advantage of a group of over 100 neurosurgeons to rerank 12
laminectomy codes from the CPT manual (63001-63047). The audience filled out a ranking sheet containing matched pairs of codes. According to the example rating sheet used by the participants, physician work was defined as “the time of the operation [times] the intensity and complexity of the procedure.” The responses were then analyzed at the University of Chicago by Prof. Ben Wright using Rasch computer software. Several codes were found to be misaligned. Procedure 63011 was seriously undervalued while 63042 was seriously overvalued. In the Spring of 1998, Dr. Florin conducted another paired comparison study; this time on 19 surgical operations in the basic reference set. He surveyed 9 general surgeons for the ranking, which ranged from CPT 99291, Critical Care, 1 hour (RVU = 4) to 48150, Pancreatectomy (RVU = 43.48). While all the codes except for Critical Care were surgical in nature, they spanned the range of surgeries across families and specialties. Again, Dr. Florin found a few codes over- or undervalued and out of order. On the other hand, many new RVU values were quite similar to original values, which is interesting given the wide range of codes and the fact that only 9 general surgeons were used for rating purposes.

The Rasch methodology involves a few steps that produce a linear scaling of codes (called Rasch measures) based on matching pairs of codes. Respondents are given one or two pages of code pairs and are asked to circle the code on each line that involves the most work (however defined). No single respondent is given all the possible pairs, and each respondent has a set of overlapping pairs that “anchor” responses across respondents. The results are then inputted into Rasch computer software that keys on the log-odds ratios of one code’s work effort versus another. The software puts the codes in rank order in terms of work effort based on the frequency of times respondents say they involve more or less work relative to other codes. It also spreads out the ranking onto a linear scale.

Once the codes are “Rasched”, Dr. Florin next converts the arbitrary psychological scale into new RVUs. The conversion is done based on an Ordinary Least Squares regression of Rasch scores onto existing RVU values. This amounts to a linear transformation of Rasch scores into RVUs, thereby maintaining the proportionality of the Rasch measures. Finally, each code’s actual current RVUs are compared to the RVUs “predicted” by its Rasch score using the linear regression line.
A Rasch Paired Comparison of Total Physician Work

Suppose the raters’ results based on paired comparisons produces a lower Rasch score for code #1 relative to code #2. Both Rasch scores would have a predicted number of RVUs based on linear regression. Then, each code’s actual RVUs would be compared with its Rasch-predicted value. If code #1 was misaligned, its Rasch-predicted RVUs might be considerably lower than its actual RVUs; lower, even, than code #2’s actual RVUs.

2.2.3 Data Requirements

Data requirements for paired comparison analyses are minimal. A small group of clinicians is required for making the paired comparisons—either at a meeting or possibly surveyed using fax or e-mail. Participants work “blind” in not knowing or discussing the way in which their colleagues are rating pairs of codes.

Worksheets would be created with 40-60 code pairs on 1-2 pages with some overlapping pairs on different worksheets to link respondent ratings. Circled codes implying more work are then used to construct a matrix reporting the number of times one code was rated higher than another and vice-versa.

2.2.4 Detailed Description of the Method

Step 1: Identify a representative small group panel of clinical experts (at least 10) familiar with a given family of codes.

Step 2: Distribute in a meeting or by mail a worksheet with 40-60 code pairs in the family and ask participants to circle the code in each pair requiring more work effort. Different definitions of work effort could be tested with the same group using separate worksheets. For example, work could be defined, first, as total work for a global 90-d service and, second, only for the intra-operative work.

Step 3: Prepare a matrix of survey results on the likelihood of one code being preferred to another for input into the Rasch software.

Step 4: Output from the software produces a Rasch yardstick graph listing each code from most to least work effort scaled, visually, on the graph. Considerable statistical output on model accuracy and reliability is also produced, such as model root mean square error and a Chi-square test of equal RVUs for all procedures.

---

Health Economics Research, Inc. The Identification of Potentially Misvalued Work RVUs: 2-33
A Rasch Paired Comparison of Total Physician Work

Step 5: Create a table listing each procedure ordered by CPT code or by current RVUs. Include a code descriptor, code number, and current MFS work units.

Step 6: Convert Rasch scores to RVUs by first regressing Rasch scores on current RVUs. Then generate a predicted (revised) RVU using each code's Rasch score.

Step 7: Transfer the new Rasch-based RVUs to your table and create a new column showing the difference between the old and new codes. Also create a Rasch-based rank order column showing misaligned codes at a glance.

Step 8: Based on the overall degree of misalignment, consider the family “aligned” and in no further need of investigation or “misaligned” and in need of further study to verify the small-group informal Rasch-based results.

Step 9: The Rasch software also indicates codes that appear far out of line with the other codes on a linear scale. These may be candidates for detailed study and may suffer from poor work definitions or simply not belong with the family of codes for some reason.

A Rasch analysis could be performed on a small or large family of codes by specialty.

2.2.5 Illustrative Example

As an example of how paired comparison analysis is performed, consider the examination of laminectomy codes, taken from Dr. Florin's research. Figure 2-1 shows a worksheet of pairs of laminectomy codes ranging from 63001 to 63047. In this example, work is defined as intra-service = surgery time x intensity. The reviewer is instructed to read across the rows one at a time and circle the code involving greater work. For example, on the first line, the reviewer is asked to rate whether 63011 involves more or less work than 63030. (The 1997 MFS work RVUs for 63011 and 63030 are 13.4 and 11.1, respectively.) The alphabetic codes at the far left signify the paired grouping. For this evaluation, code 63011 = A and 63030 = B. Several of the raters will rate combination AB, gb, etc., building up the probability matrix of relative work. Each evaluator would be given a similar, but not identical, worksheet. Some identical comparisons would appear on several worksheets to support linking the responses across raters.
### Figure 2-1

Paired Comparisons for Revaluation of Laminectomy Codes

<table>
<thead>
<tr>
<th>Group</th>
<th>Laminectomy Code</th>
<th>Laminotomy Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB</td>
<td>63011 LAMINECTOMY / SACRAL REGION</td>
<td>63030 LAMINOTOMY, ONE LEVEL FOR HERNIATED DISC, UNILATERAL, LUMBAR</td>
</tr>
<tr>
<td>gb</td>
<td>63017 LAMINECTOMY / EXPL / DECOMP, LUMBAR CORD OR CAUDA EQUINA, &gt;2 SEGS</td>
<td>63030 LAMINOTOMY, ONE LEVEL FOR HERNIATED DISC, UNILATERAL, LUMBAR</td>
</tr>
<tr>
<td>DE</td>
<td>63047 LAMINECTOMY FOR COMPLETE DECOMP, STENOSIS, LUMBAR ONE LEVEL</td>
<td>63006 LAMINECTOMY / EXPL / DECOMP CORD &amp;/OR CAUDA EQUINA, 1-2 SEGS, LUMBAR</td>
</tr>
<tr>
<td>FG</td>
<td>63001 LAMINECTOMY FOR EXPLORATION / DECOMP CORD, 1 OR 2 SEGS, CERVICAL</td>
<td>63017 LAMINECTOMY / EXPL / DECOMP, LUMBAR CORD OR CAUDA EQUINA &gt;2 SEGS</td>
</tr>
<tr>
<td>AL</td>
<td>63011 LAMINECTOMY / SACRAL REGION</td>
<td>63046 LAMINECTOMY FOR COMPLETE DECOMP, STENOSIS, CERVICAL, ONE LEVEL</td>
</tr>
<tr>
<td>CE</td>
<td>63020 LAMINECTOMY, ONE LEVEL FOR HERNIATED DISC, UNILATERAL, CERVICAL</td>
<td>63005 LAMINECTOMY / EXPL / DECOMP CORD &amp;/OR CAUDA EQUINA, 1-2 SEGS, LUMBAR</td>
</tr>
<tr>
<td>IO</td>
<td>63042 LAMINECTOMY FOR DISC, ANY LEVEL, EXTENSIVE OR RE-EXPLORAT, LUMBAR</td>
<td>63047 LAMINECTOMY FOR COMPLETE DECOMP, STENOSIS, LUMBAR, ONE LEVEL</td>
</tr>
<tr>
<td>GJ</td>
<td>63017 LAMINECTOMY / EXPL / DECOMP, LUMBAR CORD OR CAUDA EQUINA &gt;2 SEGS</td>
<td>63015 LAMINECTOMY FOR EXPLORATION / DECOMP, CERVICAL CORD, &gt;2 SEGS</td>
</tr>
</tbody>
</table>

Compare the pair of codes in each row (left to right) and circle the CPT code of the one that requires more physician work.

Physician work = the time of the operation X the intensity and complexity of the procedure.
A Rasch Paired Comparison of Total Physician Work

Figure 2-2 shows the resulting Rasch yardstick of linear measurement. This particular panel more often rated code 63040, Cervical laminotomy with decompression and/or excision of herniated disk and re-exploration, as more work than any other code in the family. Conversely, a lumbar laminotomy (63030) was rated least work relative to all the other codes. The linear measures, ranging from 020 to 158, have been derived and standardized (to code 63047) based on the frequency of respondent indications of more or less work. Based on Rasch measures, code 63047 is 5 times more work than 63030. Code 63040 is rated 58 percent more work than code 63047. The yardstick shows at a glance how similar or different the codes are in terms of perceived work. For example, 63016 and 63001 are considered essentially equal in terms of work while both are roughly double the work of 63020 and 63045.

Figure 2-3 shows the linear regression relationship between the Rasch scores and the 1997 RVUs. RVUs predicted by the Rasch scores are indicated by the white diamonds along the linear line. A companion figure (not shown) provides 95 percent confidence bands (CI’s) to the left and right of the regression line.

Table 2-2 ranks the 12 codes under consideration from least to most intra-service work according to their Rasch scores. Also listed are the actual 1997 MFS RVUs. The column headed New Work RVUs are based on predicted RVUs from a linear regression of Rasch scores on 1997 RVUs. They automatically rise from lowest to highest in accord with the Rasch ranking. The last column shows how the codes are actually ranked based on the 1997 RVUs.

From Figure 2-3 and Table 2-2 it would appear that code 63011 (A) is most out-of-line. According to the panel's evaluation, this code should be rated much higher (10th of 12) in the family in terms of work instead of second lowest. Code 63001 also appears to be somewhat undervalued. Several other codes, by contrast, appear overvalued, including 63042 and 63045. Based on the 95 percent CI’s, 3-4 procedures seem clearly misaligned, i.e., 63011, 63042, 63045, and 63016.

From this study, one could conclude that (a) the family of laminectomy codes are in need of detailed examination, and (b) code 63011 may be seriously undervalued while a couple of codes involving laminotomies with re-exploration may be overvalued. Of course, the difference in this
### A Rasch Paired Comparison of Total Physician Work

**Figure 2-2**

**RASCH Procedure Codes YARDSTICK**

*(Laminectomy)*

<table>
<thead>
<tr>
<th>RASCH LINEAR MEASURE of Amount of Physician Work</th>
<th>MOST WORK</th>
</tr>
</thead>
<tbody>
<tr>
<td>158 K</td>
<td>63040 Laminotomy::disc any level extensive or re-explor CERVICAL</td>
</tr>
<tr>
<td>146 J</td>
<td>63015 Laminectomy::explor/decomp CERVICAL cord &gt;2segs</td>
</tr>
<tr>
<td>129 A</td>
<td>63011 Laminectomy::sacral region</td>
</tr>
<tr>
<td>124 H</td>
<td>63016 Laminectomy::explor/decomp thoracic cord &gt;2segs</td>
</tr>
<tr>
<td>123 F</td>
<td>63001 Laminectomy::explor/decomp cord 1-2 segs, CERVICAL</td>
</tr>
<tr>
<td>100 D</td>
<td>63047 Laminectomy::complete decomp stenosis LUMBAR 1 level</td>
</tr>
<tr>
<td>099 E</td>
<td>63005 Laminectomy::explor/decomp cord &amp;/or cauda equina, 1-2 segs, LUMBAR</td>
</tr>
<tr>
<td>083 L</td>
<td>63042 Laminotomy::disc any level extensive or re-explor LUMBAR</td>
</tr>
<tr>
<td>076 C</td>
<td>63020 Laminotomy::1 level for herniated disc, unilateral, CERVICAL</td>
</tr>
<tr>
<td>075 I</td>
<td>63045 Laminectomy::complete decomp stenosis CERVICAL 1 level</td>
</tr>
<tr>
<td>067 G</td>
<td>63017 Laminectomy::explor/decomp LUMBAR cord or cauda equina &gt;2 segs</td>
</tr>
<tr>
<td>020 B</td>
<td>63030 Laminotomy::1 level for herniated disc, unilateral, LUMBAR</td>
</tr>
</tbody>
</table>

**LEAST WORK**

**SOURCE:** Based on Rasch Computer Software developed by J.M. Linacre & B.D. Wright, 5835 Kimbark, Chicago 60637

773-702-1596 FAX 773-834-0326

Health Economics Research, Inc. The Identification of Potentially Misvalued Work RVUs: 2-37
Figure 2-3
Paired comparisons: laminectomy
1997 Work RVUs

170
160
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10

13.4 A
14.5 F
17.43 H
13.88 E
13.57 D
13.77 G
11.4 B
16.56 L
17.77 J
K 17.56

New RVW
97 RVW
Linear (97 RVW)

1997 Medicare Work RVUs

10 11 12 13 14 15 16 17 18 19 20

Health Economics Research, Inc. The Identification of Potentially Misvalued Work RVUs: 2-38
## A Rasch Paired Comparison of Total Physician Work

### Table 2-2
Paired Comparisons for Revaluation of Codes

<table>
<thead>
<tr>
<th>Rank Order</th>
<th>CPT</th>
<th>Procedure</th>
<th>Spine: Laminectomy &amp; Laminotomy</th>
<th>1997 MFS Work RVUs</th>
<th>Sort Paired Comparison Work RVUs</th>
<th>New Work RVUs</th>
<th>1997 Rank Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>63030</td>
<td>LAMINOTOMY, ONE LEVEL FOR HERNIATED DISC, UNILATERAL, LUMBAR</td>
<td>11.1</td>
<td>20</td>
<td>11</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>63017</td>
<td>LAMINECTOMY/EXPL.DECOMP. LUMBAR CORD OR CAUDA EQUINA &gt;2 SEGS</td>
<td>14.5</td>
<td>67</td>
<td>13</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>63045</td>
<td>LAMINECTOMY/COMPLETE DECOMPR. STENOSIS, CERVICAL, ONE LEVEL</td>
<td>15.31</td>
<td>76</td>
<td>14</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>63060</td>
<td>LAMINOTOMY, ONE LEVEL FOR HERNIATED DISC, UNILATERAL, CERVICAL</td>
<td>13.77</td>
<td>76</td>
<td>14</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>L</td>
<td>63042</td>
<td>LAMINOTOMY FOR DISC, ANY LEVEL, EXTENSIVE OR RE-EXPLORAT, LUMBAR</td>
<td>16.58</td>
<td>83</td>
<td>14</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>63002</td>
<td>LAMINECTOMY/EXPL/DECOMP CORD &amp;/OR CAUDA EQUINA, 1 OR 2 SEGS, LUMBAR</td>
<td>13.88</td>
<td>99</td>
<td>15</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>63047</td>
<td>LAMINECTOMY/COMPLETE DECOMPR. STENOSIS, LUMBAR, ONE LEVEL</td>
<td>13.57</td>
<td>100</td>
<td>15</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>63001</td>
<td>LAMINECTOMY FOR EXPLORATION/DECOMP. CORD, 1 OR 2 SEGS, CERVICAL</td>
<td>14.5</td>
<td>123</td>
<td>16</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>H</td>
<td>63016</td>
<td>LAMINECTOMY FOR EXPLORATION/DECOMP. THORACIC CORD &gt;2 SEGS</td>
<td>17.43</td>
<td>124</td>
<td>16</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>63011</td>
<td>LAMINECTOMY/SACRAL REGION</td>
<td>12.4</td>
<td>129</td>
<td>16</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>J</td>
<td>63012</td>
<td>LAMINECTOMY FOR EXPLORATION/DECOMP. CERVICAL CORD &gt;2 SEGS</td>
<td>17.77</td>
<td>146</td>
<td>17</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>63040</td>
<td>LAMINOTOMY FOR DISC, ANY LEVEL, EXTENSIVE OR RE-EXPLORAT. CERVICAL</td>
<td>17.48</td>
<td>158</td>
<td>16</td>
<td>11</td>
<td></td>
</tr>
</tbody>
</table>

example may simply be due to the difference in how work was defined. For paired comparison valuation, only intra-service work was rated while the 1997 MFS RVUs naturally include pre/post work as well. In a real application, identical measures of work would be used. However, the example illustrates the technique. The results could be forwarded to the RUC for intensive review to determine if, in fact, the work of a few of the codes has changed. The suspect codes would be identified, although the RUC may not wish to share this information with any survey respondents they collect data from.

2.2.6 **Strengths and Weaknesses of Method**

*Strengths*

- Structuring small-group valuations of CPT codes using Rasch measurement techniques provides a systematic psychometric underpinning to the rating system. An extensive formal literature exists on using individual perceptions about differences in items to construct a cardinal scale ranking codes from most to least work (defined as intra-service or total).

- Rasch measurement places a minimal burden on respondents, who can generally decide which of two codes is more work within a few seconds. Filling out an entire worksheet may require 10-15 minutes.

- Unlike magnitude estimation, Rasch methods do not require clinicians to quantify the relative work of procedures.

- Small groups of clinicians would not have to meet in one place but could easily fill out their worksheets off-site and fax or e-mail them to an evaluator. Working independently would avoid the considerable time required of an entire panel of clinicians discussing and reaching consensus on the ordering and precise RVUs of codes.

- The method, while likely best suited for smaller families of codes, could also be used to structure responses across many or all of the reference set codes. As refining the reference set would require ratings from many specialists in different fields, the noninteractive Rasch worksheet approach using paired comparisons could save considerable time and resources.

- The method could be used as a first-stage approximation to a more in-depth realignment process. If applied to many families (25-50) based on off-site worksheets sent to several specialty panels, the resulting Rasch measurements may be able to systematically eliminate many families from further review as well as identifying 1 or 2 very problematic codes within a family.
Potential Weaknesses

- Rasch rank orderings will be more or less sensitive to the number and selection of clinical participants depending upon the heterogeneity of experience of the panelists. Surgeons, for example, may see systematically different patients undergoing a particular operation, e.g., benign versus malignant stomach tumor; female versus male bypass surgery. The efficient number of respondents is not known beforehand.

- Rasch rank orderings will be more or less sensitive to the equivocalness of the work definitions in a family of codes, e.g., is total work being rated or just intra-service work. This could produce unstable, inconsistent pairing by the respondents and raise questions about the existence of a linear relation among the codes.

- Practitioners may not be familiar with all the codes under study, although the method can handle blank responses. This would add error and uncertainty to some of the codes. Participants ideally would be fairly familiar with each procedure being paired and what a typical patient and practice modality would involve.

- Rasch measures would produce an alternative ordering of codes and even a new set of RVUs through regression analysis. Criteria would have to be established to guide HCFA analysts in deciding whether to leave the old codes unchanged, investigate 1 or 2 in-depth, or possibly replace with the new Rasch-based RVUs.

- Converting Rasch measures to predicted RVUs using a linear regression could compress values at the tails of the distribution. A nonlinear predicting equation may give a better fit.

- Rasch measures and resulting rank order and RVUs do not explain why some codes seem misaligned. Detailed Rasch statistics, however, can be used to identify an outlier respondent who is disproportionately influencing the anomalous results. Reasons for unusual rankings can be gleaned from the respondent, which could lead to recommendations to split codes.

- Care would have to be taken to exactly specify the type of work being rated, e.g., total work versus intra-service work.

- Given that the work component of the MFS is based on magnitude estimation, it could be inconsistent to revise one or more families of codes based on paired comparison methods.

- Rasch paired comparison does not necessarily allow for face-to-face discussion which may produce biased results.
nominated physicians to ensure their interest and availability also will be necessary as will written
certification of their appointment to a panel, once the final selections have been made.

Task 2: Develop and distribute information to the panelists

We estimate that the mail survey can be conducted during the third month or the face-to-face
meetings can be held during the third and fourth month of this project. Development of the paired
comparison worksheets can be completed during the first week of this task, regardless of data
collection mode. At the beginning of the second week, we recommend sending the mail survey
instrument to all panelists via Federal Express, following up with the panelists during the third week,
and collecting completed instruments from late responders during the fourth week. If face-to-face
meetings are held, they can commence starting the second week of this task.

Task 3: Collect and process the responses from the panelists

It is anticipated that collecting and processing the responses from the panelists should be
completed within two weeks of the survey ending or the panel meetings concluding. A matrix of
survey results on the likelihood of one code being preferred to another needs to be prepared for input
into the Rasch software.

Task 4-6: Process the survey results through the Rasch software and convert to RVUs

The output from the Rasch software must be converted to work RVUs and tables developed
displaying the current and predicted work RVUs generated from the Rasch software.
Development of tabular presentation materials of these results should be doable within a two week
time period.

Tasks 7-9: Identify misvalued services

It is anticipated that an additional four weeks would be spent identifying outlier codes from
the ratings obtained from the panelists and processed through the Rasch software. Individual codes
or small families of codes should be determined to be aligned or misaligned and referred to the RUC for review as appropriate.

**Task 10: Send to the RUC**

The last step in this method is the submission of identified services to the RUC for their review and HCFA review of any proposed new work RVUs. A specific time frame for the RUC deliberations and HCFA review is not included in Exhibit 2-3 as it can vary depending upon the number of services submitted to the RUC, the number of specialties that need to be surveyed, and availability of RUC internal resources. In Chapter 3, we provide a more general discussion of the timing of these activities within the context of the next five year review.

### 2.2.9 References


