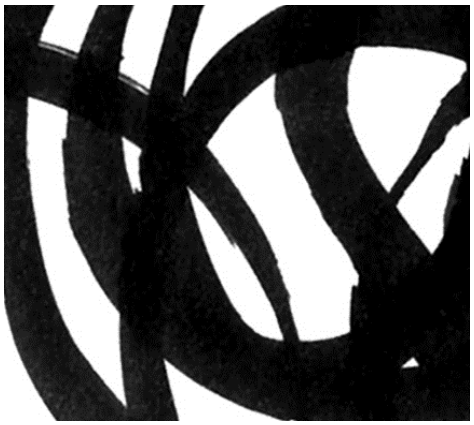
Business & Talent. Aligned.®

Right Management
ManpowerGroup



Leadership Framework 360 Tool 2012 Research Report 6 Month Analysis

EXECUTIVE SUMMARY

Prepared by:

Right Management
Paul Creighton – Psychometrician
Neil Scott - Psychometrician
Pippa Cronk – NHS LF 360 Project Manager

July 2012

75 King William Street, London, SW17 7JL
T 0207 469 6660 www.right.com

TABLE OF CONTENTS

INTRODUCTION

In June 2011, the NHS launched a new Leadership Framework. The Leadership Framework was designed to provide a consistent approach to leadership development for staff in health and care irrespective of discipline, role or function, and represents the foundation of leadership behaviour. Fundamental to its development was a desire to build on existing leadership frameworks (e.g. the LQF & MLCF) used by different staff groups and create a single overarching leadership framework for all health and care staff.

In October 2011, a new 360-degree feedback tool to support leadership development within the NHS was launched. The new tool reflects the seven domains of the Leadership Framework and is primarily aimed at leaders at stages 3 & 4 (middle managers and above). In the first 6 months since its launch, the tool has attracted over 1,300 users.

This research report was commissioned by the NHS Institute for Innovation and Improvement and covers the data collected between October 2011 and March 2012.

DETAILS OF THE SAMPLE

The database comprised 16,313 records for 1,386 participants. Demographic information was not available for 73 participants, due to these participants being 'non NHS' users.

GENDER

The gender breakdown was broadly similar to that in the 2009 LQF research report:

- 70% female
- 30% male

COUNTRY AND REGION

The Country demographic was broken down as follows:

Country	Frequency	Percent
England	1,219	92.8
Scotland	26	2.0
Wales	27	2.1
Northern Ireland	30	2.3
Isle of Man	6	0.5
Other	5	0.4

For those reporting their Country as England, the following regions were reported.

Region	Frequency	Percent
NHS North of England	382	31.3
NHS South of England	164	13.5
NHS Midlands and East	414	34.0
NHS London	200	16.4
National Organisations	49	4.0
Not Applicable	10	0.8

ORGANISATION

The organisation breakdown was as follows:

- 71.5% Provider Organisations
- 13% Commissioning Organisations
- 3.5% Strategic Health Authority
- 12% Other

ROLE

The distribution for the respondents' role is given below:

Role	Frequency	Percent
Administration	24	1.8
Allied Health Professional	100	7.6
Clinical Psychology	5	0.4
Commissioning	87	6.6
Communications	5	0.4
Dentistry	5	0.4
Facilities	10	0.8
Finance	35	2.7
Governance/Risk	23	1.8
Healthcare Science	32	2.4
Info Management and Technology	21	1.6
Medical	163	12.4
Midwifery	34	2.6
Nursing	263	20.0
Operational Management	129	9.8
Optometry	2	0.2
Other Clinical	10	0.8
Patient & Public Involvement	3	0.2
Pharmacy	41	3.1
Strategic Planning/Perf Management	24	1.8
Workforce/OD/HR	192	14.6
Other	99	7.5

LEVEL

Participants gave their levels as follows:

Level	Frequency	Percent	% by Group
Chair	21	1.6	
Non-Executive Director	0	0	
Chief Executive	2	0.2	11.8
Executive Director	29	2.2	
Director	77	5.9	
Consortia Leader	26	2.0	
Manager (Bands 8D & 9)	86	6.5	6.5
Manager (Bands 8A & 8C)	520	39.6	39.6
Manager/Team Leader (Bands 5 - 7)	339	25.8	25.8
Clinical (up to Band 6)	89	6.8	
Non-Clinical (up to Band 6)	17	1.3	8.4
Other professional (up to Band 6)	5	0.4	
Not Applicable	102	7.8	7.8

The 2009 LQF research sample had similar proportions at the most senior grades. However, there were fewer 8A/8C users in the 2009 report.

AGE

Age Group	Frequency	Percent
16 - 25	10	0.8
26 - 35	200	15.4
36 - 45	484	37.2
46 - 55	518	39.8
56 - 65	85	6.5
66+	3	0.2

The 2009 LQF age profile was broadly similar to that shown above.

ETHNICITY

19 people chose not to disclose their Ethnicity. Responses are recorded below:

Ethnicity	Frequency	Percent	2009 Percent
Asian or Asian British	102	7.9	5.6
Black or Black British	55	4.3	4.3
Chinese	3	0.2	0.4
Mixed	17	1.3	1.3
White - British	1,040	80.4	81.1
White – Other White Background	60	4.6	6.5
Other Ethnic Background	17	1.3	0.8

The ethnicity profile was only slightly different in the 2009 study, there being less Asian participants and more White participants in 2009. Ethnic Minority participants (i.e. all non-White) accounted for 15.0% in the current study. This continues the trend for increasing percentages of Ethnic Minority participants over the years.

- 2007 – 8.8%
- 2008 – 11.3%
- 2009 – 12.4%
- 2012 – 15.0%

USE OF THE RATING SCALE

A new rating scale, based on competence, was implemented in October 2011. The rating scale descriptors are given below. Numerical ratings are shown to indicate that high ratings equate with stronger performance.

Coding	Scale Point
1	Significant development needed
2	Development needed
3	Competent
4	Significant strength
5	Exemplary/Best possible
NA	Not Applicable
NO	Not Observed

SELF RATINGS VS NON-SELF RATINGS

Self-ratings were lower than non-self ratings (a trend that was found in the 2009 LQF 360 research).

- Self Mean – 3.34
- Non-Self Mean – 3.81

The following table records the percentages of all ratings that were found at each value:

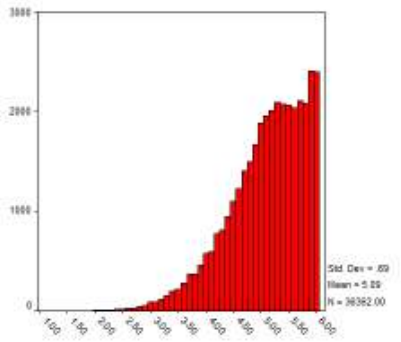
	Significant development needed	Development needed	Competent	Significant strength	Exemplary/ Best possible
Self	0.7%	13.7%	44.0%	34.0%	7.6%
Non-self	0.4%	5.6%	29.1%	41.3%	23.7%

Ratings of 'Significant development needed' are very rare but the other points on the scale are used and provide a useful range of ratings.

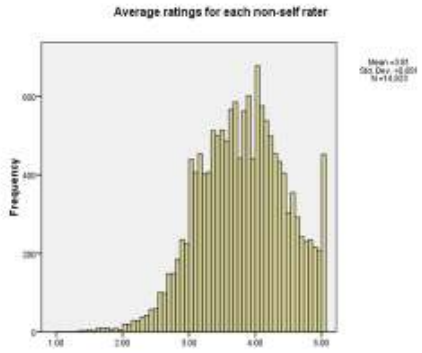
Comparative data from the old 2009 *frequency* rating scale are shown below:

	Never	Almost never	Some-times	Usually	Nearly always	Always
Self	0.1%	0.8%	8.9%	26.3%	39.8%	24.1%
Non-self	0.1%	1.0%	6.9%	16.8%	31.8%	43.4%

2009 – Average ratings for each non-self rater



2011/12 – Average ratings for each non-self rater



The 2011/2012 graph above suggests that the current scale is more effective, with many fewer raters choosing the highest available rating, thereby allowing greater discrimination between participants, particularly at the high end. The graph of the 2009 average ratings for each non-self rater also demonstrates the significant reduction of the ceiling effect in the 360 ratings (by comparison with the 2011/12 graph).

The NHS 360 scale aligns with best practice recommendations highlighted in a recent study around minimising the positive bias found in 360s. The scale has a mid-point, clearly labels all ratings and uses a development/competence scale (as opposed to an agreement scale). In addition Right Management designed the top 3 points on the 5 point scale to differentiate between high levels of effectiveness (competent, strong and truly exceptional behaviour) in order to provide greater discrimination at the top end of the scale and to reduce the positive bias that we had found in previous research. The results of this latest analysis show that this seems to be working.

RATER TYPES & AVERAGE RATINGS

There are differences between rater types in terms of how high they rate. The table below shows the average rating for each rater type.

Respondent Type	Average rating	Minimum average rating	Maximum average rating	Standard deviation of ratings
Self	3.34	2.00	5.00	0.49
Manager	3.54	1.52	5.00	0.58
Direct Report	3.91	1.05	5.00	0.67
Peer/Colleague	3.76	1.43	5.00	0.63
Other	3.89	1.30	5.00	0.65

Of all the non-self raters, Managers give the lowest average ratings, with Direct Reports and Others giving the highest average ratings.

- 11% of all ratings given by Managers stated that development was needed (i.e. were ratings below 'Competent', though, presumably, 'Competent' people still have scope for development).
- 6% of all ratings given by Direct reports were suggesting that at least some development was needed (i.e. were ratings below 'Competent').
- 6% of all ratings given by Peers/Colleagues stated that development was needed (i.e. were ratings below 'Competent').
- 5% of all ratings given by Others stated that development was needed (i.e. were ratings below 'Competent').
- 14% of all self ratings stated that development was needed (i.e. were ratings below 'Competent').

Direct Reports and Others are more likely to choose 'Not Observed'. Managers are more likely to evaluate a question as 'Not Applicable'. The number of 'Not applicable' and 'Not Observed' ratings decreases with increasing seniority (i.e. job level).

NORMATIVE DATA

Normative data (derived from average Non-self ratings) were calculated. To establish whether there may be a need for different normative data depending on level of role within the organisation, average Non-self ratings from the following groups were investigated:

- Chair & Non-Executive Directors
- Chief Executives and Executive Directors
- Consortia Leaders
- Directors & Managers (Bands 8D & 9)
- Manager (8A-C)
- Manager /Team Leader (Bands 5-7)
- Clinical/Non-Clinical/ Other prof (up to Band 6)

There was no evidence to suggest that there are meaningful differences between the proposed norm groups i.e. norms will be very similar for all groupings. One overall norm group will therefore be established, but differences will continue to be investigated in the annual research.

EFFECTS OF GENDER ON THE LF 360 RATINGS

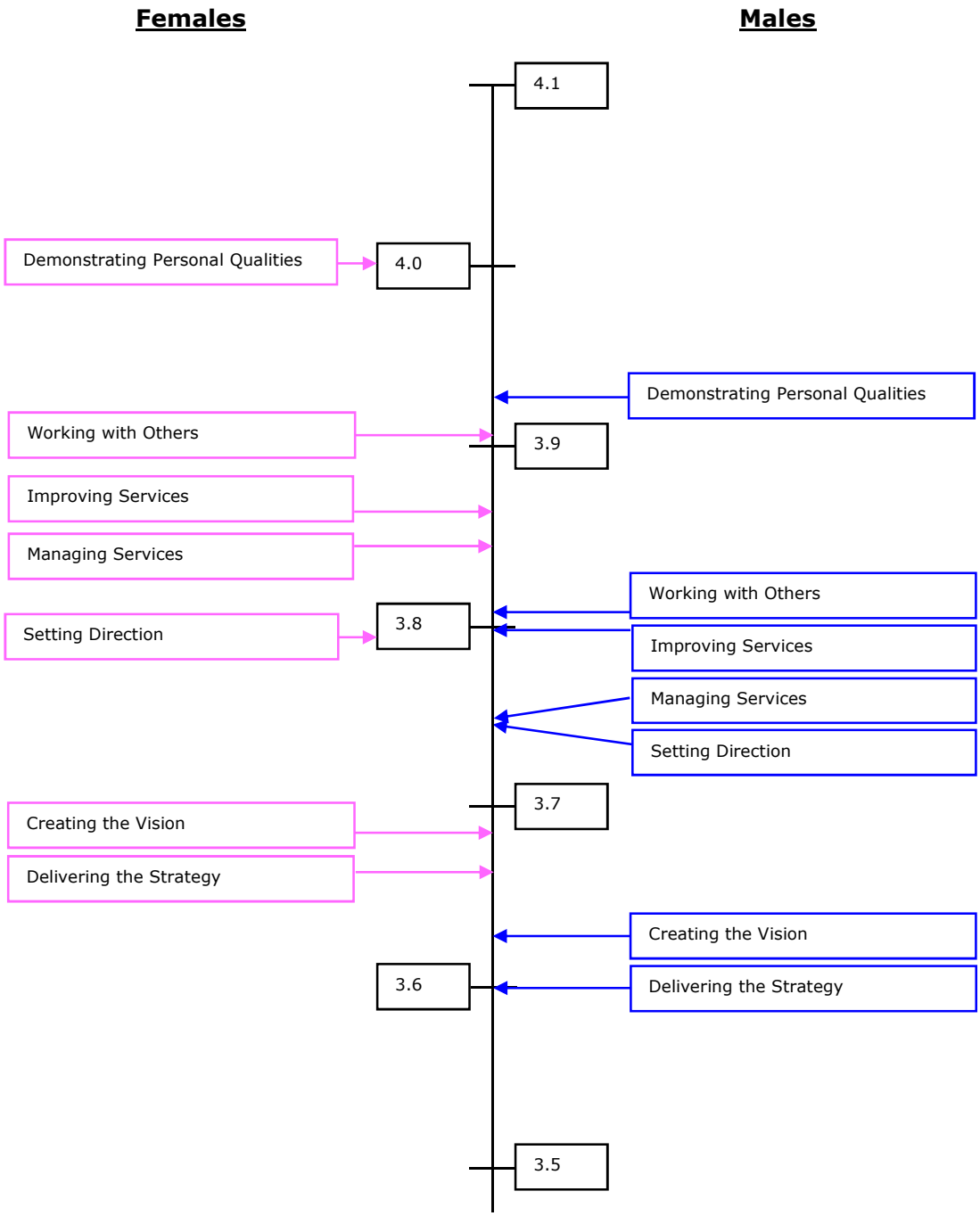
- Females are rated higher than males are rated
- Females rate themselves lower than males rate themselves

This is against a backdrop of participants rating themselves much lower than others rate them.

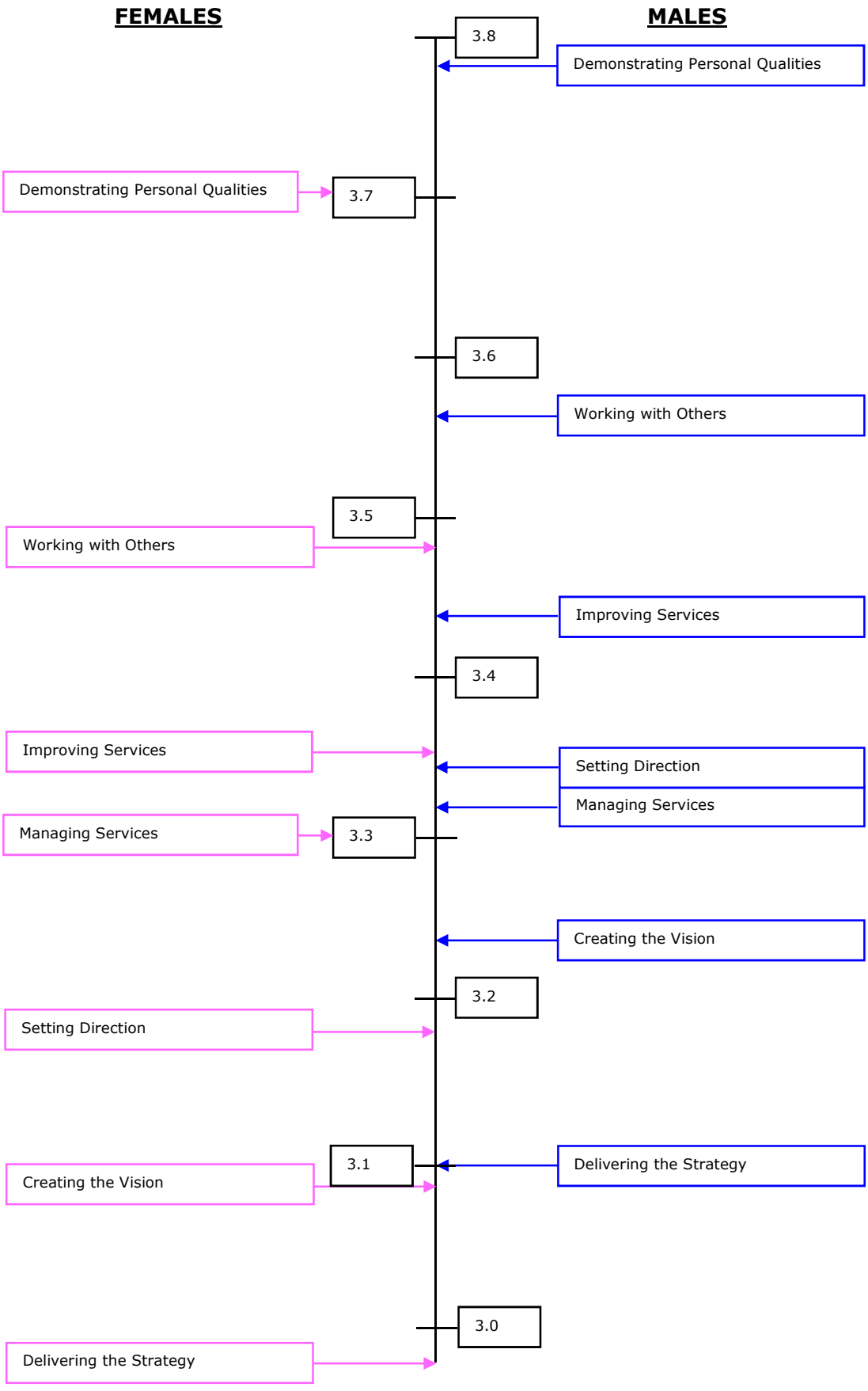
	Females	**Males**
Non-self	3.83	3.76
Self	3.31	3.41

RELATIVE RANKING OF DOMAINS

The relative Domain **non-self ratings** for Males and Females are shown in the following diagram:



The relative Domain **self ratings** for Males and Females are shown below.



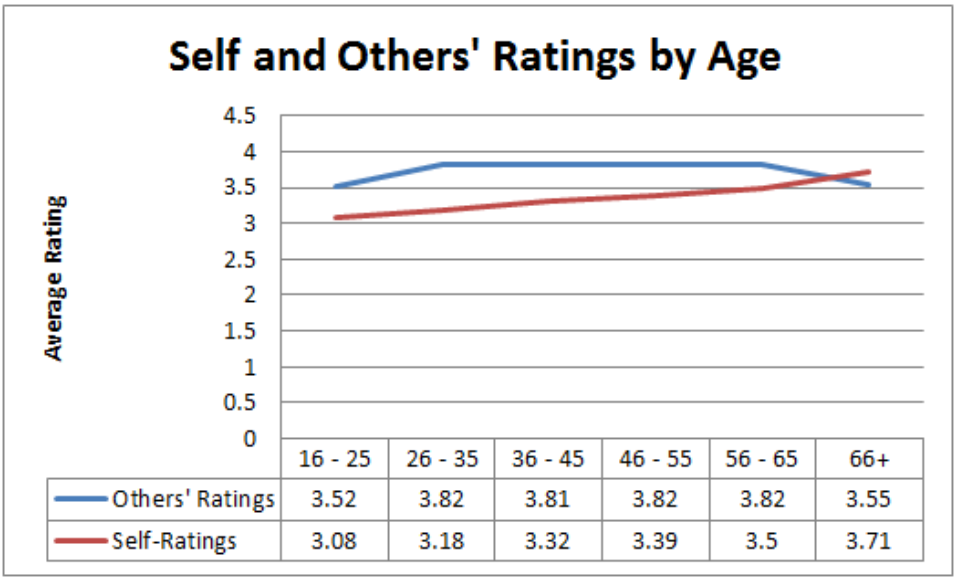
EFFECTS OF ETHNICITY ON THE LF 360 RATINGS

- Ethnic Minority participants have a mean self-rating that is higher than that of White participants
- Non-self ratings of White and Ethnic Minority participants are almost identical

	White	**Ethnic Minority**
Self	3.31	3.50
Non-self	3.81	3.80

EFFECTS OF AGE ON THE LF 360 RATINGS

Differences in the mean self and others' ratings by age are given in the table below and in the following graphic.

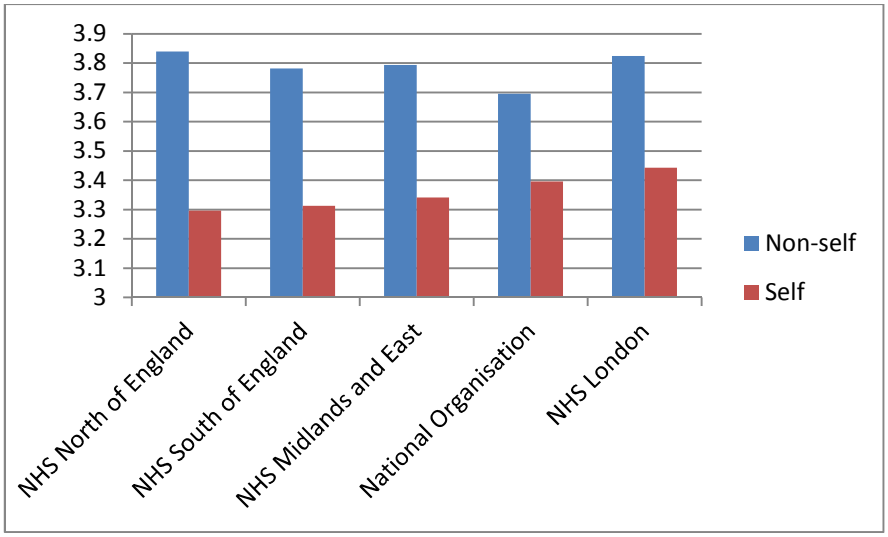


	16 - 25	26 - 35	36 - 45	46 - 55	56 - 65	66+
Others' Ratings	3.52	3.82	3.81	3.82	3.82	3.55
Self-Ratings	3.08	3.18	3.32	3.39	3.5	3.71

It can be seen that, with the exception of the youngest and oldest groups, which are both very small (10 participants in the youngest group, 3 in the oldest), non-self ratings are unaffected by age. Self-ratings, by contrast, increase steadily with age. These findings are broadly similar to those from the previous 2009 LQF research.

EFFECTS OF REGION ON THE LF 360 RATINGS

There are differences between Regions in terms of both mean self and mean non-self ratings (see graph below).



However, it seems likely that some, if not all, of these differences can be explained by differences between Regions in gender balance. We have seen that Males tend to have higher self ratings than females, and the two Regions with the highest self ratings are those with the greatest proportions of males as shown in the table below.

		Region - numeric coding					
		NHS London	NHS Midlands and East	NHS North of England	NHS South of England	National Organisation	Total
Gender	Male	80	107	107	53	21	368
		40.2%	25.9%	28.0%	32.3%	42.9%	30.5%
	Female	119	306	275	111	28	839
		59.8%	74.1%	72.0%	67.7%	57.1%	69.5%

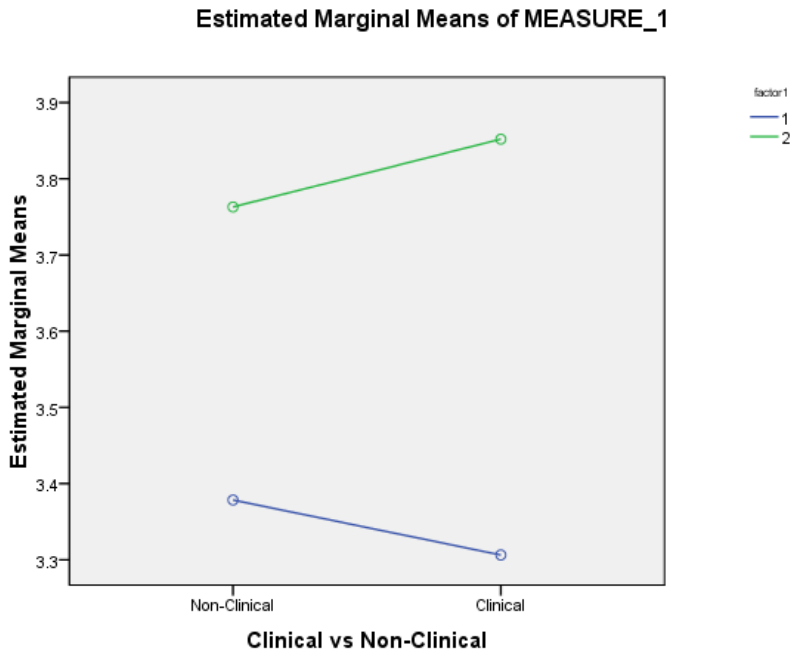
The data is not strong enough to suggest that any observed differences in ratings between the Regions are driven by the Region per se, rather than by demographic factors.

EFFECTS OF CLINICAL VS NON-CLINICAL ROLES ON THE LF 360 RATINGS

The Clinical / Non-Clinical distinction is defined by Role. The following Roles are deemed Clinical:

- Allied health professional
- Clinical psychology
- Dentistry
- Healthcare science
- Medical
- Midwifery
- Nursing
- Optometry
- Other clinical
- Pharmacy

The data shows that Clinical staff rate themselves lower than Non-Clinical staff rate themselves. Furthermore, Clinical staff are rated higher than Non-Clinical staff by others.



Key
Blue - Self Ratings
Green – Non-self Ratings

There is no difference in the proportion of Males and Females in Clinical vs. Non-Clinical staff, so the above differences cannot be explained by differential gender balance.

VALIDATION AND RELIABILITY

The purpose of validation is to demonstrate that the questionnaire measures what it is supposed to measure. In the absence of external criteria data to compare it to, (and those working in the 360 field often question whether criteria data can be found that is more valid than the questionnaire itself), validation must be an internal study. That is, the way the questions, Elements and Domains work internally and the way they relate to each other is assessed to establish whether they look the way they should if the questionnaire works, rather than producing random or faulty results.

Reliability is fundamentally concerned with the stability of the questionnaire. Conventionally this means:

- whether consistent responses would be obtained from the same person;
- whether raters rate consistently;
- whether there is consistency in what is being measured i.e. that the questions in a Domain are measuring broadly the same thing.

No person is rated twice by the same rater so we are not able to establish the first form of reliability.

Inter-rater reliability is problematic with 360 as the expectation is that different raters will give slightly different ratings and we do not have the same set of raters for each person being rated.

Finally, internal consistency can be measured but with the rider that we do not want very high correlations between questions as, whilst this is easy to obtain (by writing very similar questions) it means the breadth and richness of each element and Domain is not being measured.

As such, the validity and reliability analyses for the questionnaire are combined in an internal confirmatory study, specifically, exploring the relationships between the questions, Elements and Domains.

QUESTIONS AND ELEMENTS

There are 2 questions per Element and 4 Elements per Domain. The 8 questions for each Domain were correlated with each other. Within each Domain, the 2 questions that make up each Element should correlate more highly with its partner item than it does with the other six items in the Domain.

Only four question pairs had items that correlated more highly with questions elsewhere in the domain than with their own 'partner' element, and in call cases the difference between their partner and other items was small.

Overall, the internal consistency and construct validity of the Elements appears to be strong with most pairs of questions correlating more highly internally than with other items and all pairs comfortably exceeding the .60 target. The strength of the result may be due in part to the way the questionnaire is administered i.e. one Element at a time. The results are highly consistent with those of the Pilot Report (August 2011).

ELEMENTS AND DOMAINS

Each Element should correlate more highly with the other 3 Elements in its own Domain (summed) with other Domains. It should also have a conventionally acceptable alpha value of .60 or higher. The 'alpha' value represents the degree to which the items that make up the Elements and Domains are all measuring the same underlying attribute (i.e. the extent to which the items 'hang together'). This demonstrates that Domains are meaningful and have internal reliability and that Elements are best situated where the model says they should be situated.

There is a high level of inter-correlation between the various Elements of the questionnaire. The criteria that all alpha values should exceed .60 was easily met, the internal reliability of all Domains being very strong. Whilst five Elements correlated more highly with other Domains than their own, the differences were small. The results are highly consistent with those of the Pilot Study and suggests a stable structure.