What We Learned from Four Years of Sciencing the Crap Out of DevOps

Nicole Forsgren, PhD
Jez Humble

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Outline

How to make your data suck less
- Writing good survey questions
- Making sure the survey questions are good - with SCIENCE
- (These methods apply to your system and log data)

What we found... that we did (AND didn't) expect
Things about Continuous Delivery
Things about Management

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Not all data is created equal

Who here thinks surveys are sh*t?

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Not all data is created equal

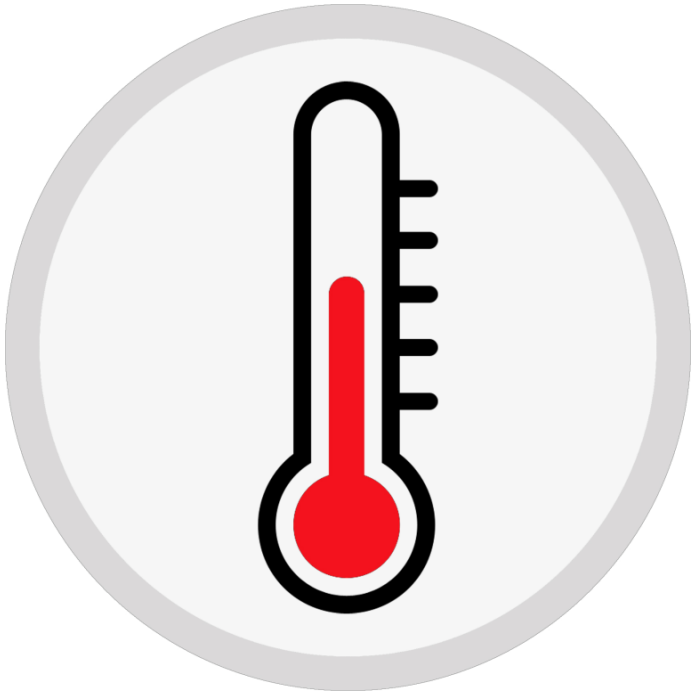
Who here thinks surveys are sh*t?

Who here LOVES the data from their log files?

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

What is a Latent Construct?



@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

We use

PSYCHOMETRICS

to make our survey data good*

**or give us a reasonable assurance that it's telling us what we think it's telling us (& some of this can also apply to your log data)*

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Psychometrics includes:

Construct creation (manual)

- When possible: use previously validated constructs
- Based on definitions and theory, carefully and precisely worded, card sorting task, pilot tested

Construct evaluation (statistics)

- Establishing Validity: discriminant and convergent
- Establishing Reliability

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Psychometrics Writing Example: Culture

- Does it matter to our study?
  - More than just intuition?
- What KIND of culture?
  - National identity and norms
  - Adaptive culture
  - Value learning (2014 study)
  - Value information flow and trust (2014 and 2015 studies -- Westrum culture)

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Psychometrics Writing Example: Culture

- Does it matter to our study?
  - More than just intuition?
- What KIND of culture?
  - National identity and norms
  - Adaptive culture
  - Value learning (2014 study)
  - Value information flow and trust (2014 and 2015 studies -- Westrum culture)

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Try writing items yourself!

Westrum typology

Pathological *Power-oriented*	Bureaucratic *Rule-oriented*	Generative *Performance-oriented*
Low cooperation	Modest cooperation	High cooperation
Messengers shot	Messengers neglected	Messengers trained
Responsibilities shirked	Narrow responsibilities	Risks are shared
Bridging discouraged	Bridging tolerated	Bridging encouraged
Failure leads to scapegoating	Failure leads to justice	Failure leads to inquiry
Novelty crushed	Novelty leads to problems	Novelty implemented

Use strong statements with clear language.

@nicolefv
@jezhumble

Westrum, R. (2004). A typology of organisational cultures. *Quality and safety in health care*, *13*(suppl 2), ii22-ii27.

Westrum Culture Items

Found to be valid & reliable

- On my team, information is actively sought.
- On my team, failures are learning opportunities, and messengers of them are not punished.
- On my team, responsibilities are shared.
- On my team, cross-functional collaboration is encouraged and rewarded.
- On my team, failure causes inquiry.
- On my team, new ideas are welcomed.

Predictive of IT Performance & Organizational Performance

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Psychometrics Analysis Example: Notification of Failure

At my organization...

- We are primarily notified of failures by reports from customers.
- We are primarily notified of failures by the NOC.
- We get failure alerts from logging and monitoring systems.
- We monitor system health based on threshold warnings (ex. CPU exceeds 100%).
- We monitor system health based on rate-of-change warnings (ex. CPU usage has increased by 25% over the last 10 minutes).

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Psychometrics Analysis Example: Notification of Failure

At my organization...

- We are primarily notified of failures by reports from customers.
- We are primarily notified of failures by the NOC.

- We get failure alerts from logging and monitoring systems.
- We monitor system health based on threshold warnings (ex. CPU exceeds 100%).
- We monitor system health based on rate-of-change warnings (ex. CPU usage has increased by 25% over the last 10 minutes).

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

More data tests!

Plus, we test to make sure the survey doesn't have other problems.

- Common method variance (CMV) (aka CMB for Bias)
- Early vs. late responders
- Survey drop-off rates and bias

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Okay **NOW** we can look at the data and how it relates to each other

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

A note about analysis methods

One of three conditions must be met:

1. Longitudinal (no, this is cross-sectional)
2. Randomized, experimental design (no, this is a non-experimental)
3. Theory-based design

When this condition was not met, only correlations were tested and reported

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

KEY FINDING:
IT Performance and its behavior

A combination of *throughput* and *stability*

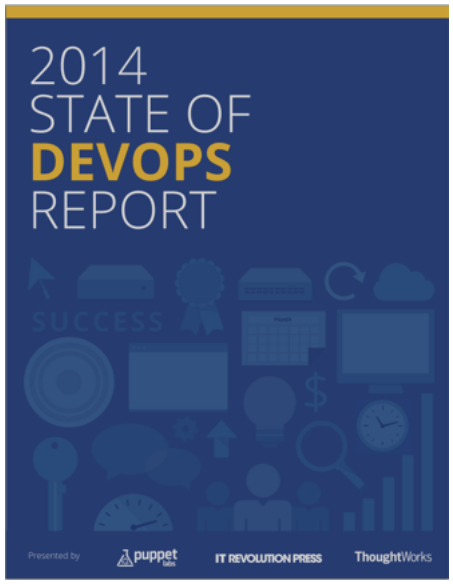
- lead time for changes
- release frequency

- time to restore service
- change fail rate

@nicolefv
@jezhumble

Forsgren, N., J. Humble (2016). "DevOps: Profiles in ITSM Performance and Contributing Factors." In the Proceedings of the Western Decision Sciences Institute (WDSI) 2016, Las Vegas, NV.

DORA
DEVOPS RESEARCH & ASSESSMENT



http://bit.ly/2014-devops-report/

KEY FINDING:
IT performance matters!

"Firms with high-performing IT organizations were twice as likely to exceed their profitability, market share and productivity goals."

IT Performance is predictive of organizational performance.

http://bit.ly/2015-devops-report/

Forsgren, N., J. Humble (2016). "The Role of Continuous Delivery in IT and Organizational Performance." In the Proceedings of the Western Decision Sciences Institute (WDSI) 2016, Las Vegas, NV.

@nicolefv
@jezhumble

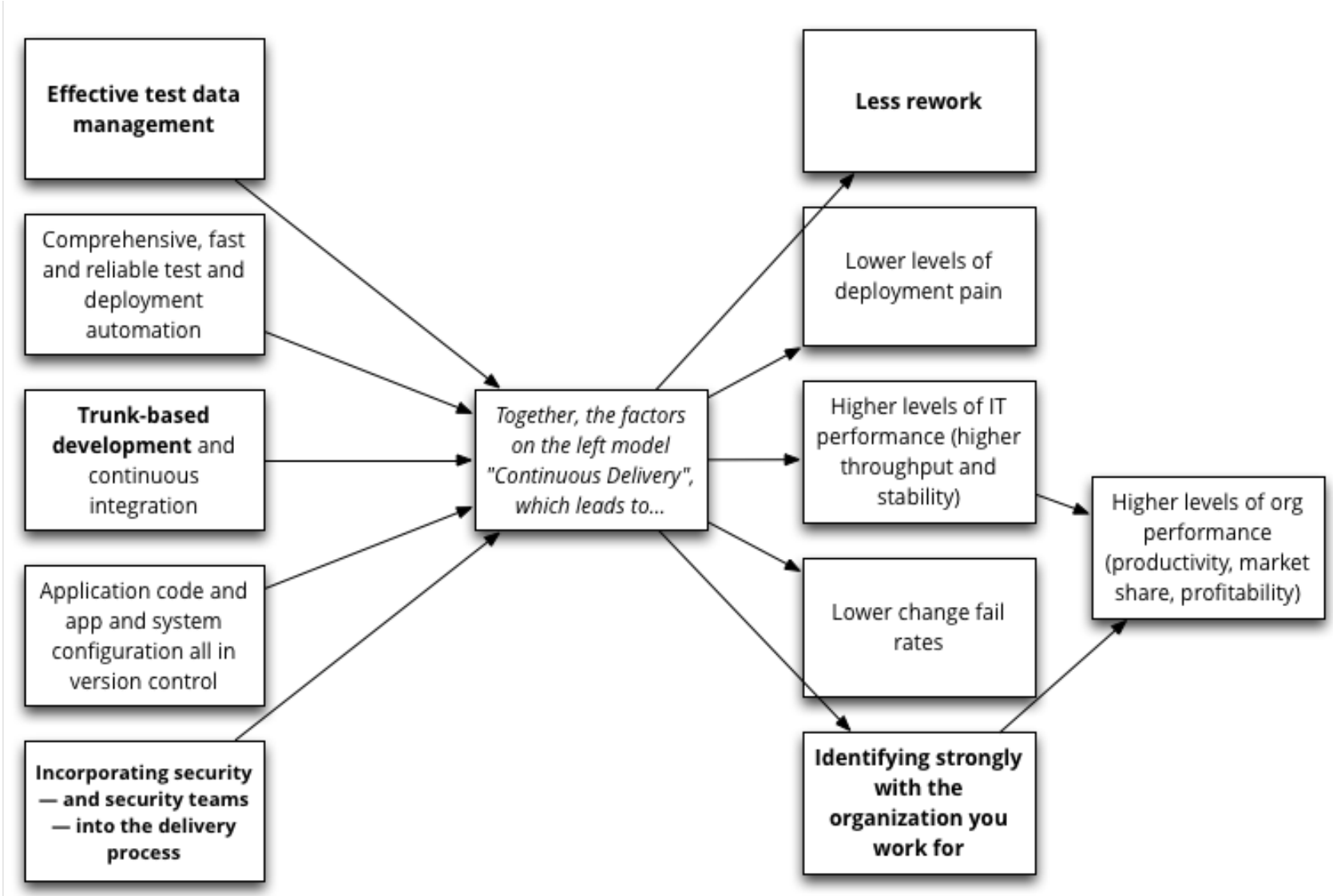
DORA
DEVOPS RESEARCH & ASSESSMENT

2016 IT Performance by Cluster

	High IT Performers	Medium IT Performers	Low IT Performers
Deployment frequency *For the primary application or service you work on, how often does your organization deploy code?*	On demand (multiple deploys per day)	Between once per week and once per month	Between once per month and once every 6 months
Lead time for changes *For the primary application or service you work on, what is your lead time for changes (i.e., how long does it take to go from code commit to code successfully running in production)?*	Less than one hour	Between one week and one month	Between one month and 6 months
Mean time to recover (MTTR) *For the primary application or service you work on, how long does it generally take to restore service when a service incident occurs (e.g., unplanned outage, service impairment)?*	Less than one hour	Less than one day	Less than one day*
Change failure rate *For the primary application or service you work on, what percentage of the changes either result in degraded service or subsequently require remediation (e.g., lead to service impairment, service outage, require a hotfix, rollback, fix forward, patch)?*	0-15%	31-45%	16-30%

* Low performers were lower on average (at a statistically significant level), but had the same median as the medium performers.

@nicolefv
@jezhumble



@nicolefv
@jezhumble

some surprises

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Which of these measure effective test practices?

Developers primarily create & maintain acceptance tests

QA primarily create & maintain acceptance tests

Primarily created & maintained by outsourced party

When automated tests pass, I'm confident the software is releasable

Test failures are likely to indicate a real defect

It's easy for developers to fix acceptance tests

Developers share a common pool of test servers to reproduce failures

Developers create on demand test environments

Developers use their own dev environments to reproduce failures

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Which of these measure effective test practices?

Developers primarily create & maintain acceptance tests

QA primarily create & maintain acceptance tests

Primarily created & maintained by outsourced party

When automated tests pass, I'm confident the software is releasable

Test failures are likely to indicate a real defect

It's easy for developers to fix acceptance tests

Developers share a common pool of test servers to reproduce failures

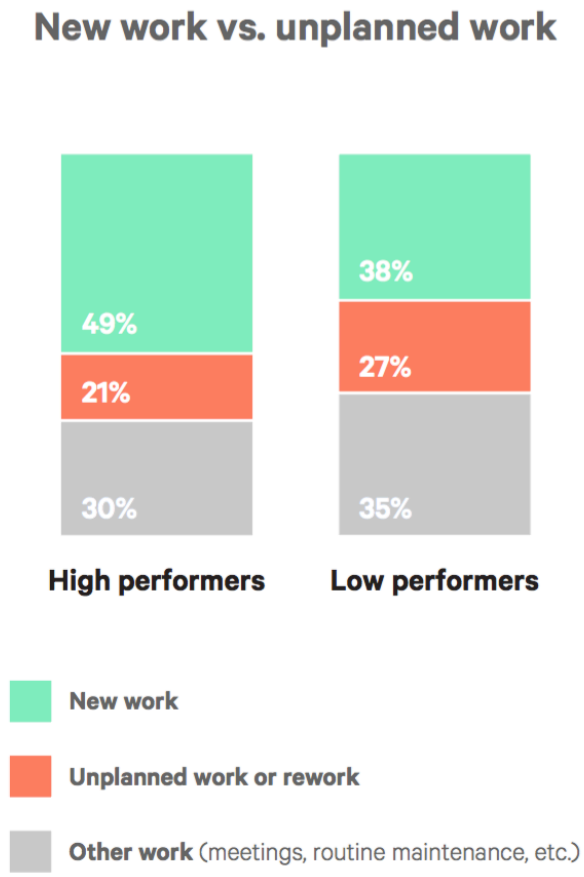
Developers create on demand test environments

Developers use their own dev environments to reproduce failures

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Key Finding: Unplanned work



High performers spend 20% more time on new work than low performers, and 22% less time on unplanned work and rework.

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Surprises with Culture

- We wanted to add additional measures of culture
  - Google study
  - Identity
  - Retain Westrum culture

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Identity & Google items

- I am glad I chose to work for this organization rather than another company.
- I talk of this organization to my friends as a great company to work for.
- I am willing to put in a great deal of effort beyond what is normally expected to help my organization to be successful.
- I find that my values and my organization's values are very similar.
- In general, the people employed by my organization are working toward the same goal.
- I feel that my organization cares about me.

Adapted from adapted from Atreyi Kankanhalli, Bernard C.Y. Tan, and Kwok-Kee Wei (2005), "Contributing Knowledge to Electronic Knowledge Repositories: An Empirical Investigation," MIS Quarterly, 29, 113-143.

@nicolefv
@jezhumble

Identity & Google items

- I am glad I chose to work for this organization rather than another company.
- I talk of this organization to my friends as a great company to work for.
- I am willing to put in a great deal of effort beyond what is normally expected to help my organization to be successful.
- I find that my values and my organization's values are very similar.
- In general, the people employed by my organization are working toward the same goal.
- I feel that my organization cares about me.

Adapted from adapted from Atreyi Kankanhalli, Bernard C.Y. Tan, and Kwok-Kee Wei (2005), "Contributing Knowledge to Electronic Knowledge Repositories: An Empirical Investigation," MIS Quarterly, 29, 113-143.

@nicolefv
@jezhumble

Westrum items

Now for management stuff

We all know managing WIP is important, right?

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Now for management stuff

We all know managing WIP is important, right?
Correlation between WIP and ITPerf is negligible

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

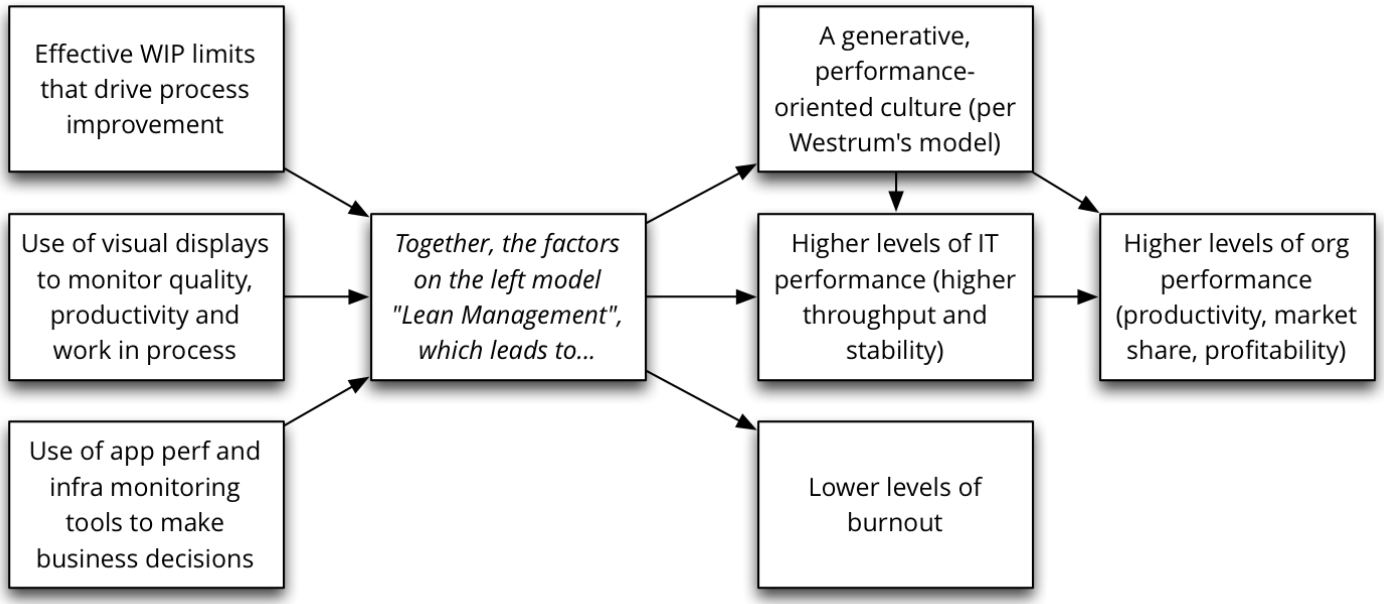
Now for management stuff

We all know managing WIP is important, right?
Correlation between WIP and ITPerf is negligible
What's going on?

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

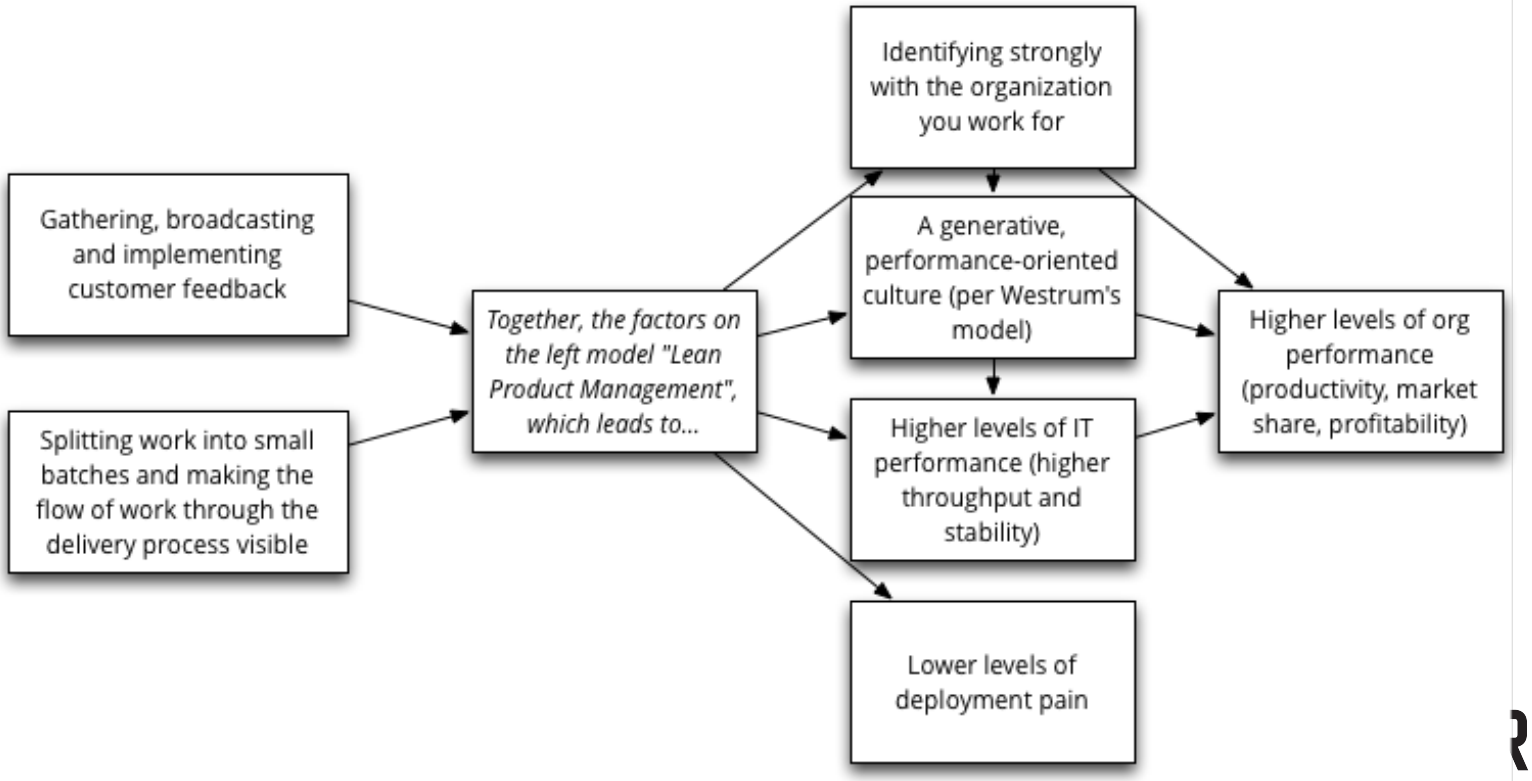
Lean management SEM



@nicolefv
@jezhumble

DEVOPS RESEARCH & ASSESSMENT

Also lean product management



@nicolefv
@jezhumble

Conclusions

- Even if you think it's obvious, TEST WITH DATA.
  - (if the results don't surprise you, you're doing it wrong)
  - (if you don't also confirm some things you expected, you're doing it wrong)
- We CAN have it all, or at least throughput AND stability.
- IT matters (but you have to do it right)
- DevOps culture & practices have a measurable impact on IT & org perf

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

For more science-ing...

Sign up for our ROI whitepaper & get peer-reviewed research
devops-research.com

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT

Thank you

devops-research.com

@nicolefv
@jezhumble

DORA
DEVOPS RESEARCH & ASSESSMENT