# 比較実験によるレビュー会議の有効性評価

久野倫義[†a)]　　中島毅[†]（正員）
松下誠[††]　　井上克郎[††]（正員）

# A Study on the Effectiveness of a Review Meeting by a Comparative Analysis

NORIYOSHI KUNO[†a)], TSUYOSHI NAKAJIMA[†], Member,
MAKOTO MATSUSHITA[††], KATSURO INOUE, Member[††]

† 三菱電機株式会社
Design Systems Engineering Center, Mitsubishi Electric Co.Ltd., 5-1-1 Ofuna, Kamakura, Kanagawa, 247-8501, Japan
†† 大阪大学
Department of Computer Science, Graduate School of Information Science and Technology, Osaka University, 1-5 Yamadaoka, Suita, Osaka, 565-0871, Japan
a) E-mail: Kuno.Noriyoshi@cb.MitsubishiElectric.co.jp

**あらまし**　Porter は実験により，レビュー会議の有効性に疑問を呈した．本論文は，開発現場で実施している条件でレビュー会議の実験を行い，レビュー会議が欠陥検出に有効な活動であることを示す．

**キーワード**　インスペクション，レビュー会議

## 1. まえがき

要求分析からコーディングに至るソフトウェア開発の上流工程では，設計文書やプログラムを対象としたレビューが主たる検証手段である．Gilb はソフトウェアインスペクションを体系化し，その中心的活動としてレビュー会議を位置付けた[1]．一方 Porter は，レビュー会議を用いるレビュー手法は，レビュー会議を用いない手法に比べて，より効果が大きいとも言えず，また少ないとも言えないと結論付けている[2]．

本論文では，開発現場で実施している条件でレビュー会議の実験を行い，ソフトウェア開発におけるレビュー会議の有効性を再評価する．

## 2. レビュー会議に関する有効性評価

実際のソフトウェア開発現場では欠陥検出のためにレビュー会議を利用する場合が多い．筆者らは，レビュー会議の欠陥検出上有効性について Porter の結論と産業界の評価にギャップがあると認識している．今回の実験では，開発現場の実施条件に合わせた実験を行い，Porter の結果と比較しその結論の再評価を行った．

### 2.1 実験条件

Porter らの Maryland 大学実験[2]と今回の実験条件を表 1 に示す．今回の実験では家電製品に関する要求分析ドキュメントを用いた．Porter の実験では 2 つのドキュメントを対象にデータを収集しているが，データ数の多い WLMS(Water Level Monitor System)を比較の対象とした．表 2 は，レビュー形態の定義である．

表 1　実験の条件
Table 1　Experiment conditions.

| 実験条件 | A: Porter 実験の値 | B: 今回の値 |
|---|---|---|
| 成果物規模/チーム員数 | 24 頁※ ／3 名 | 12 頁 ／6 名 |
| レビュー形態 | PI，DC, DD の 3 形態 | DC |
| 個人チェック時間 | 2.5 時間 | 0.5 時間 |
| レビュー会議時間 | 2.5 時間 | 0.5 時間 |

※文献[3]参照

表 2　レビュー形態
Table 2　Review-meeting methods.

| レビュー形態 | 説明 |
|---|---|
| PI:Preparation-Inspection | 個人チェックは仕様書理解に集中，レビュー会議で欠陥を検出する． |
| DC:Detection-Collection | 個人チェックで仕様書の欠陥を検出，レビュー会議で欠陥を報告，追加で新たな欠陥を検出する． |
| DD:Detection-Detection | レビュー会議を実施せず，レビューの全時間を個人チェック時間に割り当てる． |

上記 2 つの実験結果を，Votta のレビュー会議の効果尺度 Meeting-Gain（以降，欠陥検出率と呼ぶ）$R_{dm}$ を用いて評価した[3]．以下にその定義を示す．

$R_{dm} = N_{dm} / N \times 100$

$N_{dm}$:レビュー会議で検出した欠陥数

N　:検出した全欠陥数．

欠陥検出率は，個人チェックで検出できなかった欠陥をレビュー会議においてどれだけ検出できたかを評価する尺度である．

### 2.2 Porter の実験結果

Porter らは要求分析ドキュメントに対し，3 人で構成する学生とプロ開発者チームを複数編成し，レビュー形態を割り当て比較した．その結果，欠陥検出は DC 形態より DD 形態の方が多いことを示した[2]．

### 2.3 今回の実験結果

今回の実験では，13 チームが参加し要求分析ドキュメントに対しチェックリストを用い DC 形態でレビューを行った．表 3 はその実験結果である．今回の実験の $R_{dm}$ は平均 21.8%であり，文献[2]の Fig.9 から DC 形態の結果を読取り計算した 9%に比べ 12.8 ポイント高い．

表 3　今回の実験結果
Table 3　Measurement results for our experiment.

| チーム | N | $N_{dm}$ | $R_{dm}$ | チーム | N | $N_{dm}$ | $R_{dm}$ |
|---|---|---|---|---|---|---|---|
| 1 | 27 | 7 | 25.9 | 8 | 33 | 1 | 3.0 |
| 2 | 13 | 4 | 14.8 | 9 | 28 | 4 | 14.3 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 3 | 36 | 10 | 37.0 | 10 | 13 | 7 | 53.9 |
| 4 | 38 | 10 | 37.0 | 11 | 21 | 6 | 28.6 |
| 5 | 29 | 2 | 7.4 | 12 | 19 | 4 | 21.1 |
| 6 | 22 | 3 | 11.1 | 13 | 25 | 3 | 12.0 |
| 7 | 31 | 6 | 19.4 | 平均 $R_{dn}$ | | | 21.8 |

### 2.4　2 つの実験結果の統計的評価

Porter は，DD 形態と DC 形態における検出欠陥数を評価し，DD 形態が DC 形態に比べ高いことを示した．そこで，Porter の DD 形態の結果と今回の実験結果を統計的に評価した．まず 2 群の分散が同一であると仮定し，分散比の検定を行った．P 値は 0.002 となり，分散が同一であるという仮説は却下された．そこで，平均値が同等であると仮定し有意差 5%で分散比が異なる場合の t 検定を行った．その結果 P 値が 0.079 であり 2 つの平均値が同等であるという仮説を却下できず，Porter の結論とは異なる結果となった．

表 4　t 検定の結果
Table 4　Results of t-test.

| | 平均 | 観測数 | P 値 |
|---|---|---|---|
| Porter DD 形態結果（欠陥数を図より読み取った値を利用） | 29.0 | 7 | 0.079 |
| 今回の実験結果 | 21.8 | 13 | |

### 2.5　2 つの実験結果の相違に関する考察

2 つの実験結果が異なった要因を以下に示す．

(1) 個人チェック時間の設定方法

今回の測定では 12 頁の文書を用い，Porter の実験では 24 頁の文書を用いた[3]．成果物規模とレビュー参加人数，レビュー時間の関係から Porter のデータにおけるレビュー工数は，0.6 人時／頁（5 時間×3 名／24 頁）であり，今回の測定結果 0.5 人時／頁（1 時間×6 名／12 頁）とほぼ同等である．

一方，Porter 実験では，レビュー形態に関わらず Phase1 で同一の方法で個人チェックを実施し，Phase2 で，DC 形態ではレビュー会議を，DD 形態では，DC 形態のレビュー会議と同一時間をかけて個人チェックを行った．その結果，式 1 で示すピアレビュー欠陥検出率 R は，DC 形態では 0.21，DD 形態では 0.43 であった[2].

R = ピアレビューで検出した欠陥数 / 全欠陥数　（式 1）

この式は，次のように表すことができる．

$R^{xx}$ = （xx 形態の Phase1 で検出した欠陥数 ＋ xx 形態の Phase2 で検出した欠陥数）/ 全欠陥数
= $R_1^{xx}$ + $R_2^{xx}$　（式 2）

ここで xx はピアレビュー形態（DC または DD），$R^{xx}$ は xx 形態による全体のピアレビュー欠陥検出率，$R_1^{xx}$ は xx 形態による Phase1 のピアレビュー欠陥検出率，$R_2^{xx}$ は Phase2 のピアレビュー欠陥検出率である．式 3，式 4 に，DC,DD 形態のピアレビュー欠陥検出率を示す．

$R^{DC} = R_1^{DC} + R_2^{DC} = 0.21$　（式 3）
$R^{DD} = R_1^{DD} + R_2^{DD} = 0.43$　（式 4）

DD 形態の Phase1 の個人チェックの方法は DC 形態の Phase1 と同様であり，ピアレビュー欠陥検出率は等しいと考えられるので，

$R_1^{DD} = R_1^{DC}$　（式 5）
これを式 3 に代入して変形すると，$R_1^{DD} = 0.21 - R_2^{DC}$ が得られるので，
$R_1^{DD} < 0.21$　（式 6）

式 4 と式 6 から，
$R_2^{DD} = 0.43 - R_1^{DD} > 0.22$　（式 7）

成果物に内在する欠陥数が減れば，かけた労力に対して欠陥の検出量は減少するはずで，DD 形態の場合，Phase2 のピアレビュー欠陥検出率が小さくなることが予想されるが，$R_1^{DD} < R_2^{DD}$ となっている．これは Phase1 では成果物の理解に時間を要し，欠陥の検出を行うには時間が不十分だったことが原因と推測される．

今回の実験では，実験前に 2 名の開発者に依頼し，現場で個人チェックを行うのと同様に，成果物をチェックする予備実験を行っており，30 分で全体に対する指摘ができることを確認した．さらに実験後にアンケートを実施し，13 チーム×6 名全員から個人チェック時間で一通りの確認を行うことができ，時間に不足はなかったとの意見を得ている．ソフトウェア開発の現場では工期と掛けられる工数制約の中で最適なレビューを実施する．その為に個人チェックで十分に欠陥を検出してからレビュー会議を実施する．個人

チェックが不十分な状況でレビュー会議を実施しても，有効性が低いことは Gilb が主張する通りである[1].

(2) チーム編成及びレビュー会議運営方法

Porter 実験では学生とプロ開発者が参加し，今回の実験はプロ開発者のみである．この条件による差異について，文献[2]の Fig.7 の Detection Ratio のデータから DC 形態では学生チーム平均 24%，プロ平均 19.25%，DD 形態では学生チーム平均 48.6%，プロ平均 42.8%であり学生とプロ開発者チームによる相違はレビュー形態の違いに比べて無視できると判断できる．

次に，レビュー会議では，進行役，書記，読み手，レビューアの 4 つの役割が必要である[1]．Porter 実験では 3 人 1 チームであった為，各々の役割を十分果たせず，レビュー会議で新たな欠陥を検出する活動を適切に実施できなかった可能性がある．今回の筆者の実験では，進行役が議論や修正案，無発言の時間を少なくするなどの欠陥検出以外の活動を削減する改善を実施済みのレビュー会議であった為[4]，欠陥検出効率が向上していることも一因であると考える．

(3) 成果物の違いによる要因

今回の実験に用いた成果物は，家電製品の要求分析ドキュメントであり，被験者は，開発対象としてドメインの知識を有しており成果物を理解することが容易であった．一方 Porter の WLMS は，製品仕様の理解が難しかった可能性がある．

## 3. むすび

レビュー会議に関する我々の測定結果では，レビュー会議の欠陥検出率は 21.8%であり，ソフトウェア開発におけるレビュー会議の有効性は，Porter の実験と異なる結果となった．

## 文　　献


[1] T. Gilb and D. Graham: Software Inspection, Addison-Wesley, 1993.

[2] A. Porter, P. Johnson: Assessing Software Review Meetings: Results of a Comparative Analysis of Two Experimental Studies, IEEE **TSE**, Vol.23, No.3, pp.129-146, 1997.

[3] A. Porter, L. Votta and V. Basili: Comparing Detection Methods for Software Requirements Inspections: A Replicated Experiment, IEEE **TSE**, Vol.21, No.6, pp.563-575, 1995.

[4] N. Kuno, T. Nakajima, M. Matsushita, and K. Inoue: ピアレビュー有効時間比率計測によるピアレビュー会議の改善と品質改善の効果, SEC journal, No36, pp.16-23, 2014.



Abstract

Porter concludes that review meetings in inspection do not considerably contribute to effectiveness for detecting defects in target software documents. This paper shows contradictory data to the Porter's findings and our interpretation for them.



Key words

inspection, review-meeting